Improved Bounds on the Dot Product under Random Projection and Random Sign Projection

Ata Kabán
School of Computer Science
University of Birmingham
Edgbaston, UK, B15 2TT
a.kaban@cs.bham.ac.uk

ABSTRACT

Dot product is a key building block in a number of data mining algorithms from classification, regression, correlation clustering, to information retrieval and many others. When data is high dimensional, the use of random projections may serve as a universal dimensionality reduction method that provides both low distortion guarantees and computational savings. Yet, contrary to the optimal guarantees that are known on the preservation of the Euclidean distance cf. the Johnson-Lindenstrauss lemma, the existing guarantees on the dot product under random projection are loose and incomplete in the current data mining and machine learning literature. Some recent literature even suggested that the dot product may not be preserved when the angle between the original vectors is obtuse.

In this paper we provide improved bounds on the dot product under random projection that matches the optimal bounds on the Euclidean distance. As a corollary, we elucidate the impact of the angle between the original vectors on the relative distortion of the dot product under random projection, and we show that the obtuse vs. acute angles behave symmetrically in the same way. In a further corollary we make a link to sign random projection, where we generalise earlier results. Numerical simulations confirm our theoretical results. Finally we give an application of our results to bounding the generalisation error of compressive linear classifiers under the margin loss.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications—*Data mining*; I.5.1 [**Pattern Recognition**]: Models—*Statistical*

General Terms

Theory

KDD '15, August 11 - 14, 2015, Sydney, NSW, Australia

DOI: http://dx.doi.org/10.1145/2783258.2783364.

Keywords

Random projection; Dot Product Preservation; Margin Preservation; Compressive Classification

1. INTRODUCTION

Random Projection (RP) is a simple and effective method of universal dimensionality reduction that enjoys nice theoretical guarantees. It simply consists of pre-multiplying the high dimensional data with a random matrix that has the Johnson-Lindenstrauss (JL) property. Matrices with entries drawn i.i.d. from a large class of distributions, i.e. the subgaussian distributions, are known to have the JL property. Successful application areas of RP include signal processing [17, 7], theoretical computer science [14, 25, 22, 29], machine learning [4, 13, 6, 19, 24], data and text mining [9, 28], data streaming [35], and optimisation of certain functions [8]. The use of RP in conjunction with these methods gains computational efficiency for a small amount of loss in performance. Other uses of RP include cancelable biometrics [5] and privacy preserving data mining [33].

The main theoretical justification of RP is the famous Johnson-Lindenstrauss lemma (JLL) [23, 15], which guarantees that the projection of two points to a k-dimensional random subspace preserves their Euclidean distance up to a multiplicative distortion factor of $1 \pm \epsilon$ with probability higher than $1 - \exp(-k\epsilon^2/8)$. This result is known to be optimal for linear dimensionality reduction [31, 3].

Many data mining and machine learning algorithms have dot product as a key operation, for example regression and classification methods (including the SVM), and RP has been applied to these successfully in practice [9, 20, 39, 37, 34]. However, while it has been folklore that RP also preserves dot products and angles – since there is a close connection between the Euclidean distance and the dot product – the existing probability bounds on the preservation of the dot product under RP are considerably looser than those on the preservation the Euclidean distance given by the JLL. One technical reason for this is that the bounds on the dot product are typically derived by combining multiple applications of the JLL for Euclidean distances via the Bonferroni inequality (known also as the union bound) – yielding a constant factor larger than 1. It has not yet been investigated whether the use of the union bound is necessary or not for obtaining a JLL-type bound on the dot product under RP.

Beyond the above technical reason, in [32] the authors suggest a more fundamental reason why the dot product under RP may be difficult to bound as tightly as the Euclidean distance: More specifically, [32] point out that the ratio of

the standard deviation of the projected dot product and the original dot product is unbounded.

Furthermore, a recent study [39] raised some controversy as to whether the dot product would be preserved at all for obtuse angles. The authors derive a new bound on the dot product under random projection for the case when the angle between the original vectors is acute. The proof technique used in that work only applies to acute angles and the authors then suggest on the basis of numerical simulations that obtuse angles are not preserved under random projection. Unfortunately this result is highly incomplete since a particular choice of obtuse angle, as considered in the simulations, can be misleading. Indeed, as we shall see, the suggestion that obtuse angles are not preserved under RP is in fact false.

In this paper we obtain new improved bounds on the dot product under RP that take the same form as the JLL bounds for the Euclidean distance. We do this by eliminating the need to use the union bound. Instead we exploit the convexity of the exponential function within a standard Chernoff bounding argument. Our proof uses elementary techniques and works from first principles.

We also give several corollaries of this result, which make connections with previous works, including sign random projections, and we clear the confusion about the preservation of obtuse angles. Our analysis sheds light onto the issue of how the angle between the original vectors affects the relative distortion of the dot product, and reveals that the unboundedness of the coefficient of variation noted in [32] occurs only for the dot product of RPs of perpendicular vectors. This is also the only case in which we have no guarantees on the preservation of the dot product under RP. We will demonstrate numerical simulation results that confirm our theoretical findings. Finally, as an example of application of our results, we show how we can bound the generalisation error of a compressive linear classifier under the margin loss.

1.1 Background and motivation

We briefly review definitions and existing results needed in the paper.

Definition (subgaussian random variables). A 0-mean random variable X is subgaussian with parameter σ^2 if $\exists \sigma^2 > 0$ such that $\forall \lambda \in \mathcal{R}$,

$$\mathrm{E}\left\{\exp(\lambda X)\right\} \leq \exp\left\{\sigma^2\lambda^2/2\right\} \tag{1}$$

The moment generating function of X is dominated by the moment generating function of a 0-mean, σ^2 variance Gaussian that. Also, it is easy to see (by Taylor expansion) that a 0-mean subgaussian random variable has variance upper bounded by σ^2. Many useful properties of subgaussians may be found in [11].

RP uses matrices with i.i.d. subgaussian entries to linearly compress high dimensional data. The Johnson-Lindenstrauss lemma (JLL) establishes the following guarantee for the Euclidean distance of randomly projected points. The proof of this result may be found in [15, 1].

Theorem 1.1 (JLL [15, 1]). *Let $x, y \in \mathcal{R}^d$. Let $R \in \mathcal{M}_{k\times d}$, $k < d$, be a random projection matrix with entries drawn i.i.d. from a 0-mean subgaussian distribution with parameter σ^2, and let Rx, $Ry \in \mathcal{R}^k$ be the images of x, y under R. Then, $\forall \epsilon \in (0, 1)$:*

$$Pr\{\|Rx - Ry\|^2 < (1-\epsilon)\|x-y\|^2 k\sigma^2\} < \exp\left(-\frac{k\epsilon^2}{8}\right) \tag{2}$$

$$Pr\{\|Rx - Ry\|^2 > (1+\epsilon)\|x-y\|^2 k\sigma^2\} < \exp\left(-\frac{k\epsilon^2}{8}\right) \tag{3}$$

As a side note, in practice we may take the variance of the entries of R to be $1/k$ to keep the original lengths of the vectors on average. While this would simplify the notations throughout, since then $k\sigma^2 = 1$, we prefer not to commit to this choice at this point for the sake of clarity, since other choices for σ^2 are also in use in practice.

Bounding the dot product under RP is typically treated simply as a corollary of JLL, as follows (see e.g. [4] Corollary 2, [8] Lemma 3.2 , [30] Corollary 1, etc). Rewrite:

$$(Rx)^T Ry = \frac{1}{4}\left(\|R(x+y)\|^2 - \|R(x-y)\|^2\right)$$

Now, applying the JLL on both terms separately and applying the union bound yields:

$$\Pr\{(Rx)^T Ry < x^T y k\sigma^2 - \epsilon k\sigma^2 \cdot \|x\| \cdot \|y\|\} < 2\exp\left(-\frac{k\epsilon^2}{8}\right)$$

$$\Pr\{(Rx)^T Ry > x^T y k\sigma^2 + \epsilon k\sigma^2 \cdot \|x\| \cdot \|y\|\} < 2\exp\left(-\frac{k\epsilon^2}{8}\right)$$

where the factors of 2 are brought in by the use of the union bound. When we need both tails to be bounded we will of course get a factor of 4 in front of the exponential:

$$\Pr\{k\sigma^2(x^T y - \epsilon \cdot \|x\| \cdot \|y\|) \leq (Rx)^T Ry \leq k\sigma^2(x^T y + \epsilon \cdot \|x\| \cdot \|y\|)\} \geq 1 - 4\exp\left(-\frac{k\epsilon^2}{8}\right)$$

A variation on the same idea that is also in use (see e.g. in [16]), is to decompose the dot product differently:

$$(Rx)^T Ry = \frac{1}{2}\left(\|R(x-y)\|^2 - \|Rx\|^2 - \|Ry\|^2\right)$$

Then three separate applications of the JLL combined by the union bound yield the constants 3 and 6 for the one-tail and 2-tail bounds respectively.

Although for certain types of analyses the constants are not important and need not be optimal, knowing the optimal constant does matter in practice since a smaller constant implies a better guarantee than a larger one. From both theoretical and practical standpoints it is unpleasing to have weaker guarantees for the dot product under RP than we have for its induced distance – moreover since many core machine learning and data mining algorithms, such as classification and regression, rely on dot product operations, and this issue was indeed raised in [32, 35, 40].

In the light of this, and since the recent controversy about the preservation of obtuse vs. acute angles raised in [39], we give a different proof for the preservation of the dot product under RP in the next section, which eliminates the suboptimal constant factors.

2. IMPROVED BOUNDS ON THE DOT PRODUCT OF RANDOMLY PROJECTED POINTS

Here we give a new and elementary proof to bound the tails of the probability of deviation between the dot product of RP-ed vectors and the dot product of the original vectors. Our resulting bounds match the JLL bounds, which

are known to be optimal for the Euclidean distance of RP-ed vectors.

Theorem 2.1 (Dot Product under Random Projection)**.** *Let $x, y \in \mathcal{R}^d$. Let $R \in \mathcal{M}_{k\times d}$, $k < d$, be a random projection matrix having i.i.d. 0-mean subgaussian entries with parameter σ^2, and let Rx, $Ry \in \mathcal{R}^k$ be the images of x, y under R. Then, $\forall \epsilon \in (0,1)$:*

$$Pr\{(Rx)^T Ry < x^T y k\sigma^2 - \epsilon k\sigma^2 \cdot \|x\| \cdot \|y\|\} < \exp\left(-\frac{k\epsilon^2}{8}\right) \quad (4)$$

$$Pr\{(Rx)^T Ry > x^T y k\sigma^2 + \epsilon k\sigma^2 \cdot \|x\| \cdot \|y\|\} < \exp\left(-\frac{k\epsilon^2}{8}\right) \quad (5)$$

Proof. We prove the statement in eq. (4). Eq. (5) may be proved in much the same way.

Without loss of generality we replace $x \leftarrow \frac{x}{\|x\|}, y \leftarrow \frac{y}{\|y\|}$, and rewrite:

$$\begin{aligned}
&\Pr\{(Rx)^T Ry < x^T y k\sigma^2 - \epsilon k\sigma^2\} = ... \\
&\Pr\{-(Rx)^T Ry > -x^T y k\sigma^2 + \epsilon k\sigma^2\} \leq \\
&\Pr\{-x^T R^T Ry > -x^T y k\sigma^2 + \frac{1}{4}\epsilon k\sigma^2(\|x+y\|^2 + \|x-y\|^2)\}
\end{aligned}$$

The last line holds because $\frac{1}{4}(\|x+y\|^2+\|x-y\|^2) = \frac{1}{4}(2\|x\|^2+2\|y\|^2) = 1$ since $\|x\| = \|y\| = 1$.

Now, we rewrite the dot products using the parallelogram law as follows:

$$\begin{aligned}
x^T R^T Ry &= \frac{1}{4}\left(\|R(x+y)\|^2 - \|R(x-y)\|^2\right) \\
x^T y &= \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)
\end{aligned}$$

Replacing and dividing through both sides by 1/4, the probability of our interest is:

$$\begin{aligned}
&\Pr\{-\|R(x+y)\|^2 + \|R(x-y)\|^2 > ... \\
&-\|x+y\|^2 k\sigma^2 + \|x-y\|^2 k\sigma^2 + \epsilon k\sigma^2\|x+y\|^2 + \epsilon k\sigma^2\|x-y\|^2\}
\end{aligned}$$

Re-grouping, we get:

$$\begin{aligned}
\Pr\{&-\|R(x+y)\|^2 + (1-\epsilon)\|x+y\|^2 k\sigma^2 \\
&+\|R(x-y)\|^2 - (1+\epsilon)\|x-y\|^2 k\sigma^2 > 0\}
\end{aligned} \quad (6)$$

Now, using ideas from the Chernoff bounding technique, we employ the Laplace transform of both sides, so $\forall \lambda > 0$ the probability in eq. (6) equals:

$$\begin{aligned}
\Pr\{\exp[&-\lambda(\|R(x+y)\|^2 - (1-\epsilon)\|x+y\|^2 k\sigma^2) \\
&+\lambda(\|R(x-y)\|^2 - (1+\epsilon)\|x-y\|^2 k\sigma^2)] > 1\}
\end{aligned} \quad (7)$$

By Markov inequality this is upper bounded by:

$$\begin{aligned}
\mathrm{E}\{\exp[&-\lambda(\|R(x+y)\|^2 - (1-\epsilon)\|x+y\|^2 k\sigma^2) \\
&+\lambda(\|R(x-y)\|^2 - (1+\epsilon)\|x-y\|^2 k\sigma^2)]\}
\end{aligned} \quad (8)$$

Next, we introduce a new parameter: For any choice of $\alpha \in (0,1)$, the expectation in eq. (8) equals the following:

$$\begin{aligned}
\mathrm{E}\Big\{\exp\Big[&\alpha\left(-\frac{\lambda}{\alpha}(\|R(x+y)\|^2 - (1-\epsilon)\|x+y\|^2 k\sigma^2)\right) \\
&+(1-\alpha)\frac{\lambda}{1-\alpha}(\|R(x-y)\|^2 - (1+\epsilon)\|x-y\|^2 k\sigma^2)\Big]\Big\}
\end{aligned} \quad (9)$$

Using that $\exp(\cdot)$ is a convex function, the above is upper-bounded by Jensen inequality:

$$\begin{aligned}
&\alpha\mathrm{E}\left\{\exp\left[-\frac{\lambda}{\alpha}(\|R(x+y)\|^2 - (1-\epsilon)\|x+y\|^2 k\sigma^2)\right]\right\} \\
&+(1-\alpha)\mathrm{E}\left\{\exp\left[\frac{\lambda}{1-\alpha}(\|R(x-y)\|^2 - (1+\epsilon)\|x-y\|^2 k\sigma^2)\right]\right\}
\end{aligned} \quad (10)$$

where we also used the linearity of the expectation operator.

Now, since this upper bound holds for any $\lambda > 0$, we are free to choose its value to tighten the bound. In fact, observe that we are also free to chose $\alpha \in (0,1)$, which comes in handy. Denote

$$\lambda_1 := \frac{\lambda}{\alpha}; \qquad \lambda_2 := \frac{\lambda}{1-\alpha} \quad (11)$$

We may now conveniently optimise eq. (10) in λ_1 and λ_2 individually, and choose the following to approximately minimise the expression in eq. (10):

$$\lambda = \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} \quad (12)$$

This choice ensures that α is in the correct interval: Indeed, from the definition of λ_1 we have $\alpha = \frac{\lambda}{\lambda_1} = \frac{\lambda_2}{\lambda_1+\lambda_2} \in (0,1)$ as required.

Thus, in eq. (10) the probability of our interest is now bounded by a convex combination of two subexponential moment generating functions. Notice these have the same form as those that appear in the proof of the Johnson-Lindenstrauss lemma, and may be computed as follows. By Lemma 1.6. in [11], if X is subgaussian with parameter σ^2 then[1] for any $t \in [0,1)$, $\mathrm{E}[\exp(tX^2/(2\sigma^2)] \leq 1/\sqrt{1-t}$. Applying this to both terms in eq. (10), and noting that $R(x+y)$ is subgaussian with parameter $\sigma^2\|x+y\|^2$, and $R(x-y)$ is subgaussian with parameter $\sigma^2\|x-y\|^2$, (so $t := -2\lambda_1\sigma^2\|x+y\|^2$ in the first expectation and $t := 2\lambda_2\sigma^2\|x-y\|^2$ in the second one), we have that eq. (10) is upper-bounded by the following:

$$\begin{aligned}
&\alpha\left(1+2\lambda_1\sigma^2\|x+y\|^2\right)^{-k/2}\exp[\lambda_1(1-\epsilon)k\sigma^2\|x+y\|^2] + ... \\
&(1-\alpha)\left(1-2\lambda_2\sigma^2\|x-y\|^2\right)^{-k/2}\exp[-\lambda_2(1+\epsilon)k\sigma^2\|x-y\|^2]
\end{aligned} \quad (13)$$

provided that λ_1 and λ_2 are such that $\lambda_1\sigma^2\|x+y\|^2 > -\frac{1}{2}$ and $\lambda_2\sigma^2\|x-y\|^2 < \frac{1}{2}$.

Computing derivatives w.r.t. λ_1 and λ_2, equalling them to zero and solving these stationary equations yields the following optimal values after some algebra:

$$\lambda_1 = \frac{\epsilon}{(1-\epsilon)2\sigma^2\|x+y\|^2} \quad (14)$$

$$\lambda_2 = \frac{\epsilon}{(1+\epsilon)2\sigma^2\|x-y\|^2} \quad (15)$$

and it is easy to verify that they both satisfy the conditions required above.

Plugging back into eq. (13) we get after some algebra:

$$\begin{aligned}
&\alpha(1-\epsilon)^{k/2}\exp\left(\frac{k\epsilon}{2}\right) + (1-\alpha)(1+\epsilon)^{k/2}\exp\left(-\frac{k\epsilon}{2}\right) = ... \\
&\alpha\exp\left(\frac{k\epsilon}{2} + \frac{k}{2}\log(1-\epsilon)\right) + (1-\alpha)\exp\left(-\frac{k\epsilon}{2} + \frac{k}{2}\log(1+\epsilon)\right)
\end{aligned}$$

[1] This follows from eq.(1): Multiply both sides of eq.(1) by $\sigma/\sqrt{2\pi t}\exp(-\lambda^2\sigma^2/(2t))$ when $t \neq 0$ (and when $t = 0$ the stated inequality holds with equality trivially), then integrate both sides w.r.t. $\lambda \in \mathcal{R}$.

We use the following inequalities, which are easy to verify:

$$\log(1+\epsilon) \leq \epsilon - \epsilon^2/4 \quad (16)$$

$$\log(1-\epsilon) \leq -\epsilon - \frac{\epsilon^2}{2} < -\epsilon - \frac{\epsilon^2}{4} \quad (17)$$

E.g. to see eq. (16), define $f(\epsilon) = \log(1+\epsilon) - \epsilon + \epsilon^2/4$ and note that $f(0) = 0$ and $f'(\epsilon) = \frac{\epsilon(\epsilon-1)}{4(1+\epsilon)} < 0, \forall \epsilon \in (0,1)$.

Replacing, we get the claimed result:

$$\alpha \exp\left(-\frac{k\epsilon^2}{8}\right) + (1-\alpha)\exp\left(-\frac{k\epsilon^2}{8}\right) = \exp\left(-\frac{k\epsilon^2}{8}\right) \quad \square$$

An advantage of the above proof over previous ones is that it allowed us to reduce the constant in front of the exponential to 1, which now matches the JLL bounds for the Euclidean distance. We should point out that reducing this constant to 1 was previously achieved only for the special case of Gaussian RP matrices [40]. Their proof technique relies on the rotational invariance property of the Gaussian distribution, and hence it cannot be directly generalised to subgaussian RPs. Our result holds for subgaussian RPs, which represent a larger class that subsumes the Gaussian. Some examples particularly designed for computational efficiency may be found in [1].

3. COROLLARIES, DISCUSSION, AND LINKS WITH PREVIOUS RESULTS

The following immediate corollaries may be useful in applications. These will also serve us to make further comparisons and links to several previous results.

Corollary 3.1 (Relative distortion bounds)**.** *Denote by θ the angle between the vectors $x, y \in \mathcal{R}^d$. Then we have the following:*

1. *Relative distortion bound: Assume $x^T y \neq 0$. Then,*

$$Pr\left\{\left|\frac{x^T R^T R y}{x^T y} - k\sigma^2\right| > \epsilon\right\} < 2\exp\left(-\frac{k}{8(k\sigma^2)^2}\epsilon^2\cos^2(\theta)\right) \quad (18)$$

2. *Multiplicative form of relative distortion bound:*

$$Pr\{x^T R^T R y < x^T y(1-\epsilon)k\sigma^2\} < \exp\left(-\frac{k}{8}\epsilon^2\cos^2(\theta)\right) \quad (19)$$

$$Pr\{x^T R^T R y > x^T y(1+\epsilon)k\sigma^2\} < \exp\left(-\frac{k}{8}\epsilon^2\cos^2(\theta)\right) \quad (20)$$

Proof.

1. Assuming $x^T y \neq 0$ we divide both sides in eqs (4)-(5) by $x^T y$, and note that $\frac{\|x\|\cdot\|y\|}{x^T y} = \frac{1}{\cos(\theta)}$. We get:

$$\Pr\left\{\frac{x^T R^T R y}{x^T y} < k\sigma^2 - \frac{\epsilon k\sigma^2}{\cos(\theta)}\right\} < \exp\left(-\frac{k\epsilon^2}{8}\right) \quad (21)$$

$$\Pr\left\{\frac{x^T R^T R y}{x^T y} > k\sigma^2 + \frac{\epsilon k\sigma^2}{\cos(\theta)}\right\} < \exp\left(-\frac{k\epsilon^2}{8}\right) \quad (22)$$

Putting $\epsilon k\sigma^2/\cos(\theta) := \eta$ and solving for ϵ gives $\epsilon = \eta\cos(\theta)/k\sigma^2$. Replacing and renaming η to ϵ completes the proof by employing the union bound to join the upper and lower tail bounds.

2. The multiplicative form follows from eqs (21)-(22) by putting $\epsilon/\cos(\theta) := \eta$ and solving for ϵ to get $\epsilon = \eta\cos(\theta)$. Replacing and renaming η to ϵ yields the stated eqs. (19)-(20). $\square$

Corollary 3.1 tells us how the guarantees for the *relative* distortion of the dot product under RP depend on the angle between the original vectors: As we can see, there is no guarantee when the two vectors are orthogonal on each other, but for all other angles the chance of getting a relative distortion larger than a fixed tolerance ϵ decreases exponentially with $\cos^2(\theta)$ and with k.

Observe that our low distortion guarantee is symmetric around orthogonal angles – that is, contrary to the suggestion / conjecture in [39], we have symmetrically the same guarantees on the obtuse angles as we do on the acute angles. Section 4 will verify the validity of this result against empirical evidence.

It is also interesting to relate this result to the arguments and findings in [32]. In [32] the authors conjecture that it is more difficult to derive practically useful tail bounds for $x^T R^T R y$ than it is for $\|Rx - Ry\|^2$. Their argument is essentially that in the case of $x^T R^T R y$ the coefficient of variation is unbounded. In particular, [32] computed the following for the case when the entries of R are i.i.d. Gaussian with mean 0 and variance $1/k$:

$$\frac{\sqrt{\mathrm{Var}(x^T R^T R y)}}{x^T y} \geq \sqrt{\frac{2}{k}} \text{ (unbounded)} \quad (23)$$

We are interested to see what makes the above coefficient of variation unbounded. The variance is:

$$\mathrm{Var}(x^T R^T R y) = \sqrt{\frac{1}{k}(\|x\|^2\|y\|^2 + (x^T y)^2)} \quad (24)$$

So, the coefficient of variation equals the following:

$$\frac{\sqrt{\mathrm{Var}(x^T R^T R y)}}{x^T y} = \sqrt{\frac{1}{k}\left(1 + \frac{1}{\cos^2(\theta)}\right)} \quad (25)$$

We can see again, in perfect agreement with our result in Corollary 3.1, that an unbounded coefficient of variation occurs only when x and y are perpendicular. Furthermore, the coefficient of variation is symmetric around the angles $\pi/2$ and $3\pi/2$, and decreases as we move further away from these angles. Our results in Corollary 3.1 capture this behaviour while providing tail bounds rather than just a variance.

The next corollary highlights further facets of our result.

Corollary 3.2 (Margin type bounds and sign projection)**.** *Denote by θ the angle between the vectors $x, y \in \mathcal{R}^d$. Then we have the following:*

1. *Margin bound: Assume $x^T y \neq 0$. Then,*

   - *for all ρ that satisfy $\rho < x^T y k\sigma^2$ and $\rho > (\cos(\theta) - 1)\|x\|\cdot\|y\|k\sigma^2$,*

$$Pr\{x^T R^T R y < \rho\} < \exp\left(-\frac{k}{8}\left(\cos(\theta) - \frac{\rho}{\|x\|\cdot\|y\|k\sigma^2}\right)^2\right) \quad (26)$$

   - *for all ρ that satisfy $\rho > x^T y k\sigma^2$ and $\rho < (\cos(\theta) + 1)\|x\|\cdot\|y\|k\sigma^2$,*

$$Pr\{x^T R^T R y > \rho\} < \exp\left(-\frac{k}{8}\left(\frac{\rho}{\|x\|\cdot\|y\|k\sigma^2} - \cos(\theta)\right)^2\right) \quad (27)$$

2. *Dot product under sign random projection: Assume $x^T y \neq 0$. Then,*

$$Pr\left\{\frac{x^T R^T R y}{x^T y} < 0\right\} < \exp\left(-\frac{k}{8}\cos^2(\theta)\right) \quad (28)$$

Proof.

1. In eq. (4) require that $x^T yk\sigma^2 - \epsilon \cdot \|x\| \cdot \|y\| k\sigma^2 = \rho$ and solve for ϵ. This yields $\epsilon = \frac{x^T y - \rho/(k\sigma^2)}{\|x\| \cdot \|y\|} = \cos(\theta) - \frac{\rho}{\|x\| \cdot \|y\| k\sigma^2}$. Since ϵ must be in $(0, 1)$ we need to require that $\rho < x^T yk\sigma^2$ and $\rho > (\cos(\theta) - 1)\|x\| \cdot \|y\| k\sigma^2$. This proves eq. (26).

   Likewise, put $x^T yk\sigma^2 + \epsilon k\sigma^2 = \rho$ in eq. (5) to get $\epsilon = \frac{\rho/(k\sigma^2) - x^T y}{\|x\| \cdot \|y\|} = \frac{\rho}{\|x\| \cdot \|y\| k\sigma^2} - \cos(\theta)$. Again, we need $\epsilon \in (0, 1)$, which is ensured by the conditions $\rho > x^T yk\sigma^2$ and $\rho < (\cos(\theta) + 1)\|x\| \cdot \|y\| k\sigma^2$. This completes the proof of eq. (27).

2. The sign projection bounds follow as a special case of the margin bounds by taking $\rho = 0$ and considering two cases depending on the sign of $\cos(\theta)$. If $\cos(\theta) > 0$, it is easy to see that the conditions on eq. (26) are satisfied and we get:

$$\Pr\{x^T R^T Ry < 0\} < \exp\left(-\frac{k}{8}\cos^2(\theta)\right) \tag{29}$$

   If $\cos(\theta) < 0$, then we have the conditions on eq. (27), which yields:

$$\Pr\{x^T R^T Ry > 0\} < \exp\left(-\frac{k}{8}\cos^2(\theta)\right) \tag{30}$$

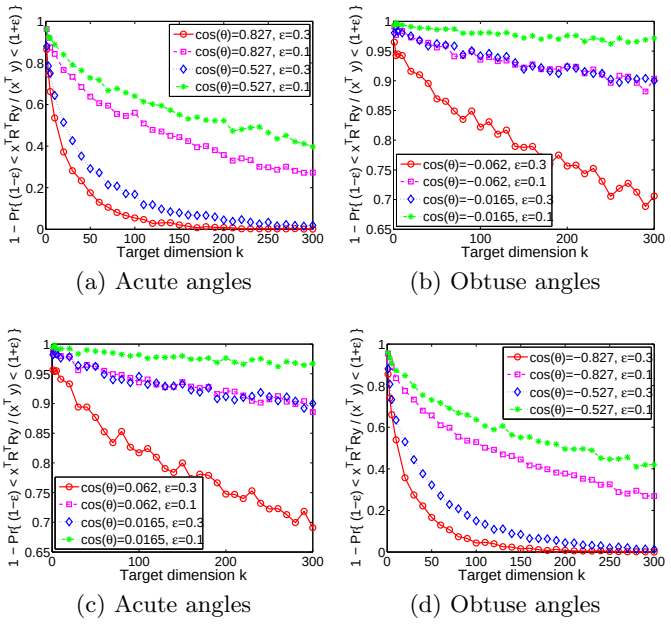
   Combining the two cases gives eq. (28). □

The form of the bound given in eq. (26) is useful e.g. in learning theory, taking $\rho > 0$, to bound the margin loss of compressive classifiers. We will develop this application in Section 5. It also serves as an intermediate step to the bound on the dot product under sign random projection, eq. (28). This bound generalises previous results on sign projection as discussed below.

The analysis of the dot product under sign projection was considered in [32], [21], and [10] for the special case when R has Gaussian entries and $k = 1$, and for this special case the exact probability of sign change is available. This was initially derived to solve a semidefinite programming problem [21] but has found use is a wide range of applications including machine learning [10], certain classes of hash functions [12], sign random projections for storage-efficient data sketches and recovery of angles between high dimensional points [32]. In [19] the sign flipping probability was computed for any $k \geq 1$, while the RP matrix R is still restricted to Gaussian. These results use the rotation invariance property of the Gaussian distribution and hence they are not directly generalisable to the subgaussian case.

In turn, our eq. (28) holds for R with any i.i.d. subgaussian entries. Also interesting to notice that, despite derived in a very different way, our probability bound has the same exponential form as the bound in the case of Gaussian R obtained in [19], with the only difference that we now have a slightly worse constant inside the exponential, i.e. 8 instead of 2 – this is a small price to pay for allowing the more general class of subgaussian RPs.

Applications of sign RP, and in particular in conjunction with dot products arose in learning theory for bounding the 0-1 loss of compressive classifiers [19]. Other uses include similarity search methods in high dimensions [29], and data classification with a better-than-chance guarantee [6]. These exploit the idea that, if x is closer to the query point than y in the data space, then there is greater chance than random guessing that this is also the case following projection onto a random line. The mentioned works used the special case $k = 1$, but one may choose a larger k to control this chance as desired. More applications of both RP and sign-RP are described in [26].



(a) Acute angles (b) Obtuse angles

(c) Acute angles (d) Obtuse angles

Figure 1: Empirical rejection probabilities of dot product preservation, i.e. the probability that the relative distortion of the dot product after RP falls outside the allowed error tolerance. The target dimension varies from 1 to the original dimension $d = 300$. (a) acute angles, (b) obtuse angles, (c) acute angles, (d) obtuse angles We see that the acute vs. obtuse nature of the angle is not the reason for the observed differences in the extent to which the dot product is preserved under RP. Note that the acute - obtuse pairs of angles that are equally distanced on the two sides of $\pi/2$ – such as (a)&(c) and (b)&(d) – have identical preservation profile.

4. NUMERICAL VALIDATION

In this section we empirically validate our theoretical results, and elucidate the issues around the preservation of obtuse angles in previous work. Specifically, our analysis predicted that the relative distortion of the dot product under RP depends on the cosine of the angle between the original vectors in a way that is symmetric about orthogonal angles, i.e. angles whose cosine is zero. So we expect to see the same behaviour for both acute and obtuse angles as long as the angle has the same distance from $\pi/2$, on either side of $\pi/2$. This is in contrast with the claim in [39] that acute angles would be preserved while obtuse angles would be not. We therefore start by replicating the experiments in [39] and demonstrating why the particular cases tested have been misleading.

We generate two vectors $x \in \mathcal{R}^d$ and $y \in \mathcal{R}^d$ of $d = 300$, at a specified angle θ between them, and generate 2000 in-

dependent instances of the RP matrix R. For the sake of concreteness and to directly compare with the results in [39] we used i.i.d. Gaussian entries $\mathcal{N}(0, 1/k)$ for these simulations. Using each of these RP matrices, we project both x and y to dimensionalities $k \in \{1, ..., 300\}$. For two different distortion tolerances, $\epsilon = 0.1$ and $\epsilon = 0.3$, we empirically estimate the probability that the relative distortion of the dot product after RP falls outside the allowed error tolerance, i.e. we compute:

$$1 - \hat{\Pr}\left\{(1-\epsilon) \leq \frac{(Rx)^T Ry}{x^T y} \leq (1+\epsilon)\right\} \tag{31}$$

We repeated this for the two acute angles $\cos(\theta) \in \{0.827, 0.527\}$ and the two obtuse angles $\cos(\theta) \in \{-0.062, -0.0165\}$ used in [39] in the first instance. These results are shown in Figure 1 (a)-(b) and these are indeed in agreement with the corresponding figures in [39].

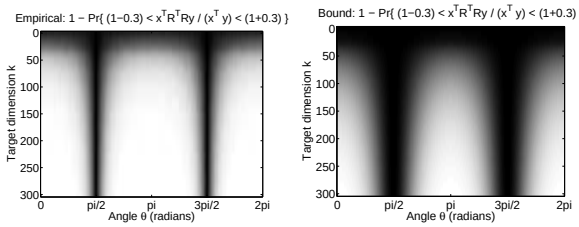
However, note that the particular acute angles tested above have a much larger absolute difference from $\pi/2$ than the particular obtuse angles tested. In fact, the latter are quite close to $\pi/2$. This is what misled the authors to the false conclusion that obtuse angles would not be preserved. To see this, we now choose the symmetrical of these angles: Two obtuse ones ($\cos(\theta) \in \{-0.827, -0.527\}$) and two acute ones ($\cos(\theta) \in \{0.062, 0.0165\}$). These results are given in Figure 1 (c)-(d). As expected, these new results with the new choices of angles suggest the exactly opposite conclusion. This is because now the chosen acute angles happen to be quite close to $\pi/2$ while the two chosen obtuse ones are much farther from $\pi/2$. In fact – as predicted by our theoretical analysis – it is most apparent from comparing the plots (a) vs. (d) in Figure 1, that this matching pair of acute and obtuse angles (i.e. that lie at equal absolute distance from $\pi/2$) do indeed exhibit the same behaviour. Likewise, the plots (b) vs. (c) of Figure 1 show another matching pair of acute & obtuse angles with identical behaviour.

In Figure 2, the leftmost column of plots depict the full picture of relative distortion estimates (cf. eq. (31)) when the angle between the original high dimensional points varies on the full range $\theta \in [0, 2\pi]$, and the target dimension varies in $k \in [0, 300]$. We repeat for $\epsilon = 0.3$ and $\epsilon = 0.1$ as before, and we also show in addition the analogous experiments for the dot product under sign random projections. For the latter we report the empirical probability of sign flipping after RP:

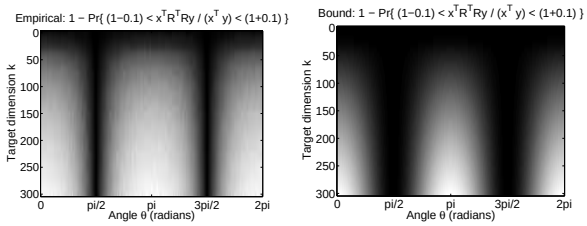
$$\hat{\Pr}\left\{\frac{(Rx)^T Ry}{x^T y} < 0\right\} \tag{32}$$

As before, all empirical probabilities are estimated from 2000 independent draws of the RP matrix R. In all of these plots, a darker grey level indicates higher probability.

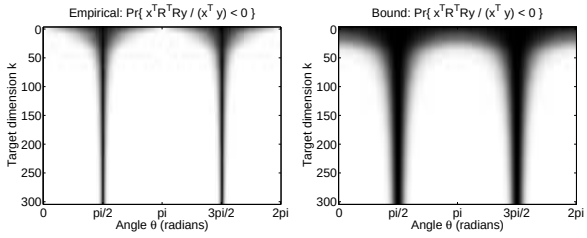
The second column of plots in Figure 2 show the corresponding theoretical bounds on these probabilities, for comparative visual inspection. For sign projection we used our generic bound for subgaussian RP from Corollary 3.2, and did not take advantage of the tighter special case for Gaussian R. In all cases it is most apparent that the true behaviour of these probabilities is well captured by the theoretical bounds. For instance, according to our theoretical analysis, the sign flipping should exhibit the same sort of dependence on the angle θ as the relative distortion of dot products under regular RP do. The angles $\theta = \pi/2$ and $\theta = 3\pi/2$ are the ones where $cos(\theta) = 0$ so these are the angles where the relative distortion can get arbitrarily large, and from our theoretical analysis the relative distortion at other angles should be symmetric around these values. This is exactly what we see in these figures. Hence we conclude that our bounds derived in the earlier sections accurately reflect the behaviour of dot product preservation, and the empirical evidence confirms our theoretical results. The extent of distortion of the dot product under RP is indeed symmetrically identical for both acute and obtuse angles.



(a) $\epsilon = 0.3$



(b) $\epsilon = 0.1$



(c) Sign projection

Figure 2: The full chart of rejection probability estimates for the tolerance values of $\epsilon = 0.1$ and $\epsilon = 0.3$, and sign flips, as the angle between the original vectors varies in $[0, 2\pi]$ and the target dimension varies from 1 to the original dimension $d = 300$. The plots in the leftmost column represent empirical probability estimates computed from 2000 independent realisations; the rightmost column of plots are the corresponding theoretical bounds. The grey levels indicate the probability of distortion: Darker means higher probability. The match between empirical behaviour and theoretical bounds is most apparent. Most importantly, all of these probabilities are symmetric around the angles of $\pi/2$ and $3\pi/2$ (i.e. orthogonal vectors before RP), which means that the preservation of the dot product is symmetrically identical for both acute and obtuse angles.

5. AN APPLICATION IN MACHINE LEARNING THEORY

There have been numerous successful practical applications of RP, as already pointed out in Section 1. In this

section we consider one of them, namely compressive classification with a norm-constrained linear classifier. SVM is a good example of this, and previous studies e.g. in [20, 39, 13] already demonstrated experiments using RP-ed high dimensional data sets. What is still lacking is a more complete theoretical understanding of the implications of using RP for dimensionality reduction before performing the classification. A theory that establishes guarantees on the performance, and that would explain the observed empirical performances is much desirable.

5.1 Margin bound on the generalisation error of compressive linear classifiers

In this section we show how we can use our results presented in the earlier sections to bound the generalisation error of norm-constrained linear classifiers that receive RP-ed data. This will be given in terms of the margin loss and the empirical Rademacher complexity of the class of classifiers under study. To do this we will apply eq (26) of our Corollary 3.2 in a special case.

Consider the hypothesis class of linear classifiers defined by a unit length parameter vector:

$$\mathcal{H} = \{x \to h(x) = w^T x : \|w\|_2 = 1\} \tag{33}$$

where $w \in \mathcal{R}^d, \|w\| = 1$. The parameters w are estimated from a training set of size N, denoted as $\mathcal{T}^N = \{(x_n, y_n)\}_{n=1}^N$, where $(x_n, y_n) \overset{\text{i.i.d}}{\sim} \mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}, \mathcal{X} \subseteq \mathcal{R}^d$.

We will work with the margin loss, which is defined as the following:

$$\ell_\rho(u) = \begin{cases} 0 & \text{if } \rho \leq u \\ 1 - u/\rho & \text{if } u \in [0, \rho] \\ 1 & \text{if } u \leq 0 \end{cases} \tag{34}$$

The generalisation error (or risk) of a learner $h \in \mathcal{H}$ is defined as $\mathrm{E}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$, and the empirical margin error of h is $\frac{1}{N}\sum_{n=1}^N \ell_\rho(h(x_n)y_n)$.

We are interested in the case when d is large and N not proportionately so. This is the case in domains where obtaining labelled examples is relatively expensive. Denote the RP matrix $R \in R^{k\times d}$, $k < d$, with entries R_{ij} drawn i.i.d. from a subgaussian distribution. The classifier only receives the randomly projected data $\mathcal{T}_R^N = \{(Rx_n, y_n)\}_{n=1}^N$.

In the reduced k-dimensional space we have analogous definitions, and we will use a subscript to refer to the reduced space. So the hypothesis class in the reduced space is

$$\mathcal{H}_R = \{x \to h_R(Rx) = w_R^T Rx : \|w_R\|_2 = \sqrt{k\sigma^2}\} \tag{35}$$

where $w_R \in R^k$ are the parameters, which are estimated from $\mathcal{T}_R^N$ by minimising the empirical margin error: $\hat{h}_R = \arg\min_{h_R \in \mathcal{H}_R} \frac{1}{N}\sum_{n=1}^N \ell_\rho(h_R(Rx_n), y_n)$.

We take the parameter of the entries of R to be $\sigma^2 = 1/k$ so the scale of the original data geometry remains the same after RP on average. Thus, $k\sigma^2 = 1$.

The quantity of our interest is the generalisation error of $\hat{h}_R$ as a random function of both $\mathcal{T}^N$, and R – that is,

$$\mathrm{E}_{(x,y)\sim\mathcal{D}}\left[\hat{h}_R(Rx) \neq y\right] \tag{36}$$

5.1.1 Generalisation bound

We will make use of the empirical Rademacher complexity of $\mathcal{H}_R$, which is defined as

$\hat{\mathcal{R}}_N(\mathcal{H}_R) = \mathrm{E}_\gamma[\sup_{h_R\in\mathcal{H}_R} \frac{1}{N}\sum_{n=1}^N \gamma_n h_R(Rx_n)]$, where $\gamma = (\gamma_1, ..., \gamma_N)$ and γ_n takes values in $\{-1, 1\}$ with equal probability. Here we used that $\mathcal{H}_R$ is closed under negation i.e. $\mathcal{H}_R = -\mathcal{H}_R$ so that no absolute value is needed in the expression of the Rademacher complexity.

Theorem 5.1. *Let R by a $k \times d, k < d$ matrix with i.i.d. 0-mean subgaussian entries with parameter $1/k$, and a compressed training set $\mathcal{T}_R^N = \{(Rx_n, y_n)\}_{n=1}^N$, where (x_n, y_n) are drawn i.i.d. from some distribution $\mathcal{D}$. For any $\delta \in (0, 1)$, the following holds with probability at least $1 - 3\delta$ for the empirical minimiser of the margin loss in the RP space, uniformly any margin parameter $\rho \in (0, 1)$:*

$$E_{(x,y)\sim\mathcal{D}}\left[\hat{h}_R(Rx) \neq y\right] \leq \min_{h\in\mathcal{H}} \left\{ \frac{1}{N}\sum_{n=1}^N 1\left(h(x_n)y_n < \rho\right) + S_k ... \right.$$

$$\left. + \sqrt{3\log(1/\delta)}\sqrt{S_k} \right\} + \frac{4}{\rho}\cdot\frac{1}{\sqrt{N}}\sqrt{\left(1 + \sqrt{\frac{8\log(1/\delta)}{k}}\right) Tr\left(\frac{XX^T}{N}\right)}$$

$$+ \sqrt{\frac{\log\log_2(2/\rho)}{N}} + 3\sqrt{\frac{\log(4/\delta)}{2N}}$$

where θ_n is the angle between the parameter vector of h and the vector $x_n y_n$. The function $1(\cdot)$ takes value 1 if its argument is true and 0 otherwise. X is an $N \times d$ matrix that holds the input points, and $S_k = ...$

$$\frac{1}{N}\sum_{n=1}^N 1\left(h(x_n)y_n \geq \rho\right) \left(\exp\left\{ -\frac{k}{8}\left(\cos(\theta_n) - \frac{\rho\sqrt{1 + \sqrt{\frac{8\log(1/\delta)}{k}}}}{\|x_n\|} \right)_+^2 \right\} + \delta \right)$$

and $(\cdot)_+ = \max\{\cdot, 0\}$.

Proof.

Fixing R at first, we start from an existing margin bound that holds for any $h_R \in \mathcal{H}_R$, uniformly for any margin $\rho \in (0, 1)$ [36] (Theorem 4.5), and which we apply in the reduced space:

$$\mathrm{E}_{(x,y)\sim\mathcal{D}}\left[\hat{h}_R(Rx) \neq y\right] \leq \frac{1}{N}\sum_{n=1}^N \ell_\rho(\hat{h}_R^T(Rx_n)y_n)$$

$$+ \frac{4}{\rho}\hat{\mathcal{R}}_N(\mathcal{H}_R) + \sqrt{\frac{\log\log_2(2/\rho)}{N}} + 3\sqrt{\frac{\log(4/\delta)}{2N}} \tag{37}$$

Now, for the empirical risk minimiser of the margin loss we have that:

$$\frac{1}{N}\sum_{n=1}^N \ell_\rho(\hat{h}_R^T(Rx_n)y_n) \leq \frac{1}{N}\sum_{n=1}^N \ell_\rho\left(\frac{(Rw)^T}{\|Rw\|}Rx_n y_n\right) \tag{38}$$

$$\leq \frac{1}{N}\sum_{n=1}^N 1\left(\frac{w^T R^T Rx_n y_n}{\|Rw\|} < \rho\right) \tag{39}$$

for any choice of $h \in \mathcal{H}$ defined by parameter vector w, because $\frac{Rw}{\|Rw\|} \in \mathcal{H}_R$, and eq. (39) is due to an upper bound on the margin loss.

We add and subtract $\frac{1}{N}\sum_{n=1}^N 1(w^T x_n y_n < \rho)$, and notice that:

$$\frac{1}{N}\sum_{n=1}^N 1\left(\frac{w^T R^T Rx_n y_n}{\|Rw\|} < \rho\right) - \frac{1}{N}\sum_{n=1}^N 1(w^T x_n y_n < \rho)$$

$$\leq \frac{1}{N}\sum_{n=1}^N 1\left(\frac{w^T R^T Rx_n y_n}{\|Rw\|} < \rho\right)\cdot 1\left(w^T x_n y_n \geq \rho\right) \tag{40}$$

Now, we exploit the fact that the right hand side in eq. (40) is a random variable that takes values in the bounded interval $[0, 1]$. Hence we can apply a Chernoff bound (Theorem

1.1 in [18]) to get the following upper bound:

$$\leq \frac{1}{N}\sum_{n=1}^{N} 1\left(w^T x_n y_n \geq \rho\right) \Pr_R\left\{\frac{w^T R^T R x_n y_n}{\|Rw\|} < \rho\right\} \ldots$$
$$+\sqrt{3\log(1/\delta)}\sqrt{\frac{1}{N}\sum_{n=1}^{N} 1\left(w^T x_n y_n \geq \rho\right) \Pr_R\left\{\frac{w^T R^T R x_n y_n}{\|Rw\|} < \rho\right\}} \quad (41)$$

w.p. $1-\delta$. From eq. (41) we see that it is enough to bound $\Pr_R\left\{\frac{w^T R^T R x_n y_n}{\|Rw\|} < \rho\right\}$ when $w^T x_n y_n \geq \rho$. We will do this by applying our Corollary 3.2 to the vectors w and $x_n y_n$. First, rewrite for any $\epsilon > 0$:

$$\Pr_R\left\{\frac{w^T R^T R x_n y_n}{\|Rw\|} < \rho\right\} \leq \ldots$$
$$\Pr_R\left\{\left(w^T R^T R x_n y_n < \rho\sqrt{1+\epsilon}\|w\|\right) \vee \left(\|Rw\| > \sqrt{1+\epsilon}\|w\|\right)\right\} \quad (42)$$

This is easy to see as the statement in the probability on the l.h.s. implies the one on the r.h.s. (the negation of the second statement implies the negation of the first).

Now, eq. (42) is upper bounded using the union bound:

$$\leq \Pr_R\left\{w^T R^T R x_n y_n < \rho\sqrt{1+\epsilon}\|w\|\right\} + \Pr_R\left\{\|Rw\| > \sqrt{1+\epsilon}\|w\|\right\} \quad (43)$$

We apply our Corollary 3.2 eq. (26) to the first term, and the Johnson-Lindenstrauss lemma to the second term. Since we have $w^T x_n y_n \geq \rho$, and $\|w\| = 1$, and the first condition needed to apply eq.(26) is $w^T x_n y_n > \rho\sqrt{1+\epsilon}\|w\|$ we will use the function $(\cdot)_+ = \max\{\cdot, 0\}$. The second condition is trivially satisfied because $\rho > 0$. Thus, we get that eq. (43) is upper bounded by the following:

$$\exp\left\{-\frac{k}{8}\left(\cos(\theta_n) - \frac{\rho\sqrt{1+\epsilon}}{\|x_n\|}\right)_+^2\right\} + \exp\left\{-\frac{k}{8}\epsilon^2\right\} \quad (44)$$

We may chose to have the last term equal to δ and express ϵ as a function of δ i.e. $\epsilon = \sqrt{\frac{8\log(1/\delta)}{k}}$.

Replacing this into eq. (41) completes the bound on the empirical margin loss term of eq. (37).

Finally, $\hat{\mathcal{R}}_N(\mathcal{H}_R)$ can be bounded by exploiting that $\|\hat{h}_R\|_2 \leq 1$, in the same way as in [27] with a modification that allows us efficient bounding w.h.p. with respect to the random draws of R:

$$\begin{aligned}
\hat{\mathcal{R}}_N(\mathcal{H}_R) &= \mathrm{E}_\gamma\Big[\sup_{h_R \in \mathcal{H}_R} \frac{1}{N}\sum_{n=1}^{N}\gamma_n h_R(Rx_n)\Big] \\
&= \mathrm{E}_\gamma\Big[\sup_{w_R:\|w_R\|_2\leq 1} \frac{1}{N} w_R \sum_{n=1}^{N}\gamma_n Rx_n\Big] \\
&\leq \frac{1}{N}\mathrm{E}_\gamma\Big\|\sum_{n=1}^{N}\gamma_n Rx_n\Big\|_2 \quad \text{(by Cauchy-Schwartz)} \\
&\leq \frac{1}{N}\sqrt{\mathrm{E}_\gamma\Big\|\sum_{n=1}^{N}\gamma_n Rx_n\Big\|_2^2} \quad \text{(by Jensen ineq.)} \\
&= \frac{1}{N}\sqrt{\sum_{n=1}^{N}\|Rx_n\|_2^2} \quad \text{(independence of } \gamma_1, \ldots \gamma_N\text{)} \\
&\leq \frac{1}{\sqrt{N}}\sqrt{(1+\epsilon)\frac{1}{N}\sum_{n=1}^{N}\|x_n\|_2^2} \qquad (45)
\end{aligned}$$

with probability $1 - \exp(-k\epsilon^2/8)$. Eq. (45) follows by a Chernoff bound for dependent variables [38]. Rearranging, we get that w.p. $1-\delta$,

$$\hat{\mathcal{R}}_N(\mathcal{H}_R) \leq \frac{1}{\sqrt{N}}\sqrt{\left(1+\sqrt{\frac{8\log(1/\delta)}{k}}\right)\mathrm{Tr}\left(\frac{XX^T}{N}\right)} \quad (46)$$

where X is the matrix that holds the input points.

Combining the three probability bounds, i.e. eq. (46), eq. (44) and eq. (37) that each hold w.p. $1-\delta$, the union bound gives the statement of the theorem. □

5.1.2 Discussion

Some comments are now in order. There has been previous work bounding the error of compressive classifiers based on the preservation of a margin [4, 6, 13, 37]. The works of [4] and [6] concern the case of classes that are linearly separable by a margin. The results in [13] and [37] do not make this assumption but they assume that the data has a sparse representation. The result in [37] is specific to the SVM.

In contrast, we do not assume class separability, and do not assume a sparse representation – in fact we do not make any assumption on the data. Margin bounds hold true irrespective of class separability, and so does our bound in Theorem 5.1. Of course, as all margin bounds, our bound is tightest when most points do have a large margin, since this is when the empirical error term (the first term on the r.h.s.) is small. Also note that the margins of all points participate in the bound. In particular, $\cos(\theta_n)$ is the normalised margin of the n-th point, and the terms of S_k decay exponentially with this quantity. The average of unnormalised margins also features in the first term on the r.h.s., i.e. the empirical error term. Overall, the bound may be seen as a margin distribution bound for compressive linear classifiers.

It is insightful to look at what the obtained bound tells us. Notice that, apart from the S_k terms, the r.h.s. of eq.(5.1) has the same flavour as the margin bound in data space – with the margin ρ balancing between the empirical margin (first term) and the Rademacher complexity (fourth term). The latter shrinks with $\rho/\sqrt{\mathrm{Tr}(XX^T/N)}$ (i.e. margin divided by diameter) just like the empirical Rademacher complexity of the uncompressed function class $\mathcal{H}$ does. To make this link more precise, we can rewrite eq. (46) as the following:

$$\hat{\mathcal{R}}_N(\mathcal{H}_R) \leq \sqrt{1+\sqrt{\frac{8\log(1/\delta)}{k}}} \cdot \hat{\mathcal{R}}_N(\mathcal{H}) \quad (47)$$

We see this is no larger than $(1+\epsilon)$ times the empirical Rademacher complexity of the original $\mathcal{H}$, where $\epsilon = \sqrt{\frac{8\log(1/\delta)}{k}}$.

Now, S_k is a penalty we get for working in the RP space: It penalises the good points to some extent – i.e. those points that have margin at least ρ – by how close their margin actually is from (an $\sqrt{1+\epsilon}$ multiple of) ρ. These penalties shrink exponentially with k, and the points whose margin is in the small interval $[\rho, \rho\sqrt{1+\epsilon}]$ will get penalty 1 just as the points that didn't have a margin ρ. Overall, this is the term that reveals some influential characteristics of the data distribution that permit the classification task to be solved well and efficiently in a random subspace by the considered function class. In the present case this is governed by the distribution of margins of the uncompressed train-

ing points, namely most points should have a margin larger than $\rho\sqrt{1+\epsilon}$.

It might be interesting to note also that the bound does not depend on d but only on the margin parameter ρ. Due to the norm constraint on the parameter vectors w_R and w, this is also the case for the dataspace margin bound – the Rademacher complexity does not depend on d as the VC dimension would. So, for the class of classifiers analysed here we do not get a reduction of the complexity term by doing RP – in contrast to the VC bound based analysis in [19]. Such reduction would only occur in the unregularised linear function class analysed in [19], and could be more pronounced in the case of peculiar data distributions where VC bounds are tight.

Figure 3 illustrates the predictive behaviour of our bound on the UCI data set of Advert classification, in comparison with the alternative VC-type bound in [19], and the actual performance estimated on holdout sets using SVM with default settings and 30 random splits (in proportion 2/3 training & 1/3 testing) of the data. We used $\delta = 0.1$ and $\rho = 0.05$. This data set was previously used in [20], it has $d = 1554$ features and $N = 3279$ points. We standardised the data first, and scaled it to $\max_{n\in\{1,\ldots,N\}} \|x_n\| = 1$.

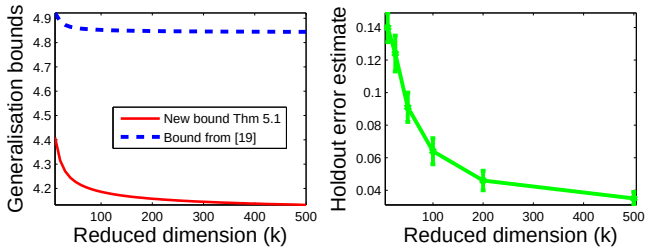


Figure 3: Illustration of the theoretical error bound of Theorem 5.1 in comparison with the VC-type bound in [19] and the actual performance of SVM on holdout sets, using the Advert classification data set. The error bars on the latter plot represent one standard deviation over 30 independent random train-test splits of the data. The gain in the tightness of our new bound is most apparent, as well as the agreement between the theoretical error bound (which predicts performance based on training errors only) and the actual holdout estimates of the performance.

As expected at this sample size, our new bound is tighter than the VC bound in [19] and both bounds predict well the trend in the holdout errors. The gain in tightness is due to the tighter complexity term that is able to take advantage of the norm constraint on the parameter vector and does not grow with the input data dimension (k), whereas the VC complexity grows with $\sqrt{k}$. This is of course offset by the magnitudes of S_k that are obviously larger than the corresponding flip probabilities in [19] (which correspond to the case $\rho = 0$) – in line with the spirit of margin bounds. We have not investigated whether the expression of S_k could be improved, e.g. if the union bound in eq.(43) in the proof could be avoided as well, this remains for future work.

6. CONCLUSIONS

In this paper we proved new bounds on the dot product under random projection that take the same form as the optimal bounds on the Euclidean distance in the Johnson-Lindenstrauss lemma. This is pleasing for at least two reasons: From the conceptual point of view, it is the inner product that defines a norm, in particular the dot product defines the Euclidean norm – hence a guarantee that holds on the latter was quite natural to expect to hold also on the former. From the practical point of view, the dot product is ubiquitous in algorithms for data mining and machine learning. The use of random projections to speed up such algorithms or to perform privacy preserving data mining without much distortion in performance can now be better justified. Previous guarantees on the preservation of the dot product under random projections have been loose and incomplete, and left room for controversy. Our results shed light on these issues and give the precise way in which the relative distortion of the dot product under random projection depends on the angle between the original vectors. This dependence turned out to have an intimate relationship with the notion of margin in the context of generalisation theory. We also obtained connections with sign random projections, generalising earlier results. We demonstrated numerical simulations that confirm the validity of our theoretical results. Finally we provided an application of our results in learning theory, where we derived a new margin distribution type generalisation error bound for compressive linear classifiers under the margin loss.

Our proof technique applies to any independent-entry subgaussian random projection matrix, which includes sparse and bit-flipping variants [1]. An interesting question for future work would be to investigate whether it could be adapted to other Fast JL transforms such as those in [2].

7. REFERENCES

[1] D. Achlioptas. Database-friendly Random Projections: Johnson-Lindenstrauss with Binary Coins. Journal of Computer and System Sciences, 66(4):671–687, 2003.

[2] N. Ailon, B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. SIAM J. Comput., 39(1), 302-322.

[3] N. Alon. Problems and results in extremal combinatorics, Part I. Discrete Math, 273:31-53, 2003.

[4] R.I. Arriaga, and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In 40th Annual Symposium on Foundations of Computer Science (FOCS), pp. 616–623, 1999.

[5] D. Arpit, I. Nwogu, G. Srivastava, V. Govindaraju. An analysis of random projections in cancelable biometrics. arXiv:1401.4489 [cs.CV], 2014.

[6] M.F. Balcan, A. Blum, S. Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings, Machine Learning 65 (1), 79-94, 2006.

[7] R.G. Baraniuk, M. Davenport, R.A. DeVore, and M.B. Wakin. A simple proof of the restricted isometry property for random matrices. Constructive Approximation 28, no. 3, pp. 253-263, 2008.

[8] A. Barvinok. Integration and optimization of multivariate polynomials by restriction onto a random subspace. Foundations of Computational Mathematics, 7, pp. 229-244, 2007.

[9] E. Bingham and H. Mannila. Random projection in dimensionality reduction: Applications to image and

text data. In Knowledge Discovery and Data Mining (KDD), pp. 245–250, ACM Press, 2001.

[10] A. Blum. Random projection, margins, kernels, and feature-selection. In SLSFS 2005, ed. Saunders et al. No. 3940 in LNCS. pp. 55-68, 2006.

[11] V.V. Buldygin, Y.V. Kozachenko. Metric characterization of random variables and random processes. American Mathematical Society, 2000.

[12] M. Charikar. Similarity estimation techniques from rounding algorithms. In Proceedings of the thirty-fourth annual ACM symposium on Theory of computing. ACM. pp. 380-388, 2002.

[13] R. Calderbank, S. Jafarpour, and R. Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical report, Rice University, 2009.

[14] S. Dasgupta. Learning mixtures of Gaussians. In Annual Symposium on Foundations of Computer Science (FOCS), vol. 40, pp 634-644, 1999.

[15] S. Dasgupta, and A. Gupta. An elementary proof of the Johnson–Lindenstrauss Lemma. Random Structures & Algorithms, 22:60–65, 2002.

[16] M.A. Davenport, M.B. Wakin and R.G. Baraniuk. Detection and estimation with compressive measurements. Technical Report TREE 0610, Rice University, January 2007

[17] D.L. Donoho. Compressed sensing, IEEE Trans. Information Theory 52, no. 4, pp. 1289-1306, 2006.

[18] D.P. Dubhashi, A. Panconesi. Concentration of measure for the analysis of randomized algorithms. Cambridge University Press, 2012.

[19] R.J. Durrant, A. Kabán. Sharp generalization error bounds for randomly-projected classifiers. 30th International Conference on Machine Learning (ICML), Journal of Machine Learning Research-Proceedings Track 28(3):693-701, 2013.

[20] D. Fradkin, D. Madigan. Experiments with random projections for machine learning. Proc. 19-th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 522-529, 2003.

[21] M. Goemans, D. Williamson. Improved approximation algorithms for maximum cut and satisability problems using semidenite programming, Journal of the ACM (JACM) 42, 1145, 1995.

[22] P. Indyk, R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. Proceedings of the 30-th annual ACM Symposium on Theory of computing, New York, NY, USA, pp. 604-613, 1998.

[23] W.B.Johnson, J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. Conference in Modern Analysis and Probability (New Haven, Conn., 1982), Contemporary Mathematics 26, Providence, RI: American Mathematical Society, pp. 189-206, 1984.

[24] A. Kabán. New bounds on compressive linear least squares regression. Proc. of the 17-th International Conference on Artificial Intelligence and Statistics (AISTATS), Journal of Machine Learning Research-Proceedings Track, vol. 33, pp. 448-456, 2014.

[25] A.T. Kalai, A. Moitra, G. Valiant. Disentangling Gaussians. Communications of the ACM 55, no. 2, pp. 113-120, 2012.

[26] Z.S. Karnin, Y. Rabani, A. Shpilka. Explicit dimension reduction and its applications. SIAM Journal on Computing, 41(1), pp. 219-249, 2012.

[27] S.M. Kakade, K. Sridharan, A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. Advances in Neural Information Processing Systems 21, pp. 793–800, 2008.

[28] S. Kaski. Dimensionality reduction by random mapping: fast similarity computation for clustering. In Proceedings of the 1998 IEEE International Joint Conference on Neural Networks, vol. 1, pp 413-418, 1998.

[29] J. Kleinberg. Two algorithms for nearest-neighbor search in high dimensions. In Proceedings of the twenty-ninth annual ACM Symposium on Theory of Computing (STOC 1997), ACM. p. 608, 1997.

[30] S. Krishnan, C. Bhattacharyya, and R. Hariharan. A randomized algorithm for large scale support vector learning. In Advances in 20th Neural Information Processing Systems. 793–800, 2008.

[31] K.G. Larsen, J. Nelson. The Johnson-Lindenstrauss lemma is optimal for linear dimensionality reduction, arXiv preprint arXiv:1411.2404, 2014.

[32] P. Li, T. Hastie, K. Church. Improving random projections using marginal information. In Proc. Conference on Learning Theory (COLT) 4005, 635-649, 2006.

[33] K. Liu, H. Kargupta, J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering 18(1), pp. 92-106, 2006.

[34] S. Marukatat. Classification with sign random projections. PRICAI 2014: Trends in Artificial Intelligence Lecture Notes in Computer Science Volume 8862, pp 708-719, 2014.

[35] A.K. Menon, G.V.A. Pham, S. Chawla, A. Viglas. An incremental data-stream sketch using sparse random projections. In Proceedings of the 2007 SIAM conference on data mining (SDM), Minnesota, USA.

[36] M. Mohri, A. Rostamizadeh, A. Talwalkar. Foundations of machine learning. The MIT Press, 2012.

[37] S. Paul, C. Boutsidis, M. Magdon-Ismail, P. Drineas. Random projections for linear support vector machines. ACM Trans. Knowl. Discov. Data 8, 4, Article 22, 2014.

[38] A. Siegel. Toward a usable theory of Chernoff bounds for heterogeneous and partially dependent random variables. Technical Report, New York Univ., 1995.

[39] Q. Shi, C. Shen, R. Hill, A. Hengel. Is margin preserved after random projection? Proceedings of the 29th International Conference on Machine Learning (ICML), pp. 591–598, 2012.

[40] E. Skubalska-Rafajlowicz. Neural networks with sigmoidal activation functions: dimension reduction using normal random projection. Nonlinear Analysis 71, pp. e1255-e1263, 2009.