Nonlinear Analysis 71 (2009) e1255–e1263

Contents lists available at ScienceDirect

Nonlinear Analysis

journal homepage: www.elsevier.com/locate/na

Neural networks with sigmoidal activation functions—dimension reduction using normal random projection

Ewa Skubalska-Rafajłowicz

Wrocław University of Technology, Institute of Computer Engineering, Control and Robotics, Wybrzeże Wyspiańskiego 27, 50 370 Wrocław, Poland

ARTICLE INFO

Keywords:
Dimension reduction
Normal random projection
Inner product
Neural network

ABSTRACT

The dimensionality and the amount of data that needs to be processed when intensive data streams are observed grows rapidly, together with the development of sensors arrays, CCD and CMOS cameras and other devices. The aim of this paper is to analyze an approach to dimensionality reduction as a first stage of the multi-layer feed-forward neural networks with sigmoidal activation functions. The lower bound on dimension of the output of dimensionality reduction layer is established. Tail bounds for inner products – similar to the known Johnson–Lindenstrauss results, concerning squared distance preservation – are given.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

The idea presented in this paper is to add a new dimensionality reduction layer to neural networks with sigmoidal activation functions. In contrast to multi-layer nets, in which all layers are nonlinear, the proposed layer is linear in input variables. Furthermore, weights of this layer are chosen in a random way, instead of using a kind of learning process. At first glance it can be surprising, but recent results on random projections (see the bibliography cited in Section 3) provide us with tools for dimensionality reduction, which retain (with a prescribed accuracy and probability) not only the Euclidean distances between pairs of projected points, but also the inner product. Using random projections as the first layer (very similar to neuronal random projection [1,2]), we are able to reduce the dimensionality of input data, deteriorating their metric relationships only slightly. This approach is closely related to machine learning theory [3], where random projections are considered as a robust concept learning tool which can be employed to reduce the dimensionality of examples, independent of the concept class[1]. Their approach is based on the observation that random projection preserves approximately the distances between a pair of points (Johnson–Lindenstrauss lemma [4]).

The dimensionality reduction layer consists of inner products of input vectors and weight vectors — each weight vector is used to form a new coordinate value. Here we examine the properties of the dimensionality reduction layer in more depth, since we have found out how normal random projection preserves the inner product of vectors. More generally, the question we consider in this paper is whether a neural network with sigmoidal activation functions and a dimensionality reduction layer is able to approximate the corresponding neural network with the non-reduced input dimension.

Although the observation that the inner product is preserved (approximately) by random projections, is made by other authors [5,6], in this paper we have shown the useful tail bounds for concentration of inner products around the mean (see Lemma 1 in Section 4).

The paper is organized as follows. In the next section we formulate the main theorem, which establishes the lower bound on the dimension of the output of the dimensionality reduction layer, assuming that the rest of the network structure remains unchanged (i.e. the number of neurons in the hidden layer and the weights of the output layer) and the weights in

E-mail address: ewa.skubalska-rafajlowicz@pwr.wroc.pl.

0362-546X/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.na.2009.01.124

the input layer are also projected, using the same projection as used in the dimensionality reduction layer for projecting the input vector. This lower bound ensures that the outputs of both networks differs only slightly (by $\varepsilon > 0$) for n different input vectors with probability $1-\delta$. The dimension after projection depends only on the number of input vectors, the number of neurons in the hidden layer, the proscribed accuracy ε and the probability of failing δ the approximate equivalence of the networks under consideration.

The proof of the dimensionality reduction theorem is based on properties of normal random projections of inner products. In Section 3 we provide basic facts concerning random projections, which are used for stating the result on the accuracy of a random projection layer in comparison to nets without dimensionality reduction. Then, in Section 4, we present the lemma formulating tail bounds for inner products similar to the Johnson–Lindenstrauss results concerning squared distance preservation [7]. The proof of the lemma is given in the Appendix.

2. Properties of a neural network with sigmoidal activation functions and dimensionality reduction layer

We turn our attention to a very well-known neural network with sigmoidal activation functions [8,9].

Let

$$y(x) = v_0 + \sum_{i=1}^{M} v_i f\left(\sum_{j=1}^{d} w_{i(j)} x_j\right), \quad x \in X \subset R^d, \tag{1}$$

be a feed-forward neural network with sigmoidal activation function $f : R \to [-1, 1]$ – nondecreasing, Lipschitz continuous function, such that $f(t) \to 1$ if $t \to \infty$ and $f(t) \to -1$ if $t \to -\infty$. Furthermore, we assume that X is a compact set in R^d, as a prerequisite to the universal approximation properties of sigmoidal neural networks [10,11].

Further, define the corresponding neural network with a dimensionality reduction layer as:

$$y_S(x) = v_0 + \sum_{i=1}^{M} v_i f\left(\frac{1}{k}\sum_{j=1}^{k} \hat{w}_{i(j)} \hat{x}_j\right), \quad x \in R^d \tag{2}$$

where $\hat{x}_j = \sum_{l=1}^{d} s_{jl} x_l$ $(j = 1, \ldots, k)$ and $\hat{w}_{i(j)} = \sum_{l=1}^{d} s_{jl} w_{i(l)}$ $(j = 1, \ldots, k\ i = 1, \ldots, M)$ are a projection of vector $x \in R^d$ and $w_i \in R^d$ onto R^k using $S \in R^{k\times d}$ – a random matrix whose entries are samples from a normal random variable, i.e., $s_{ij} \sim N(0, 1)$ which are independent and identically distributed (i.i.d.).

We prove the following theorem.

Theorem 1. *Define $R = C_L \max_i |v_i| \|w_i\| \max_{x\in X} \|x\|$, where $C_L > 0$ is Lipschitz constant depending on an activation function f. Choose $\varepsilon \in (0, 1)$ and $\delta > 0$. If the dimensionality after reduction k is selected in such a way that*

$$k \geq \frac{\log 2 + \log M + \log n - \log \delta}{c_1\varepsilon^2 - c_2\varepsilon^3}, \tag{3}$$

where $c_1 = \frac{1}{4R^2}$ and $c_2 = \frac{1}{6R^3}$, then the probability that

$$|y(x_{(i)}) - y_S(x_{(i)})| > \varepsilon, \quad i = 1, \ldots, n$$

is smaller than $\delta > 0$, where $x_{(i)} \in X$, $i = 1, \ldots, n$, is a given set of n points.

Proof. Note, that using the union bound, for any x and for any $\varepsilon > 0$, we obtain

$$\begin{aligned}
Pr\{|y(x) - y_S(x)| > \varepsilon\} &= Pr\left\{\left|\sum_{i=1}^{M} v_i \left[f\left(\frac{1}{k}\sum_{j=1}^{k} \hat{w}_{i(j)}\hat{x}_j\right) - f(w_i^{\mathrm{T}} x)\right]\right| > \varepsilon\right\} \\
&\leq Pr\left\{\sum_{i=1}^{M} |v_i| \left|f\left(\frac{1}{k}\sum_{j=1}^{k} \hat{w}_{i(j)}\hat{x}_j\right) - f(w_i^{\mathrm{T}} x)\right| > \varepsilon\right\} \\
&\leq \sum_{i=1}^{M} Pr\left\{|v_i| \left|f\left(\frac{1}{k}\sum_{j=1}^{k} \hat{w}_{i(j)}\hat{x}_j\right) - f(w_i^{\mathrm{T}} x)\right| > \varepsilon\right\} \\
&\leq M \max_i Pr\left\{\left|f\left(\frac{1}{k}\sum_{j=1}^{k} \hat{w}_{i(j)}\hat{x}_j\right) - f(w_i^{\mathrm{T}} x)\right| > \frac{\varepsilon}{|v_i|}\right\}.
\end{aligned}$$

Furthermore, we have assumed that f is Lipschitz continuous

$$|f(t + a) - f(t)| \leq C_L|a|, \quad t, a \in R,$$

where $C_L > 0$ is the constant depending on f. Note, that for activation functions such as *tanh* or logistic function $(1-\exp -t)/(1+\exp -t)$ we obtain $C_L = 1$ and $C_L = 1/2$, respectively.

Thus, we obtain

$$\left| f\left(\frac{1}{k}\sum_{j=1}^{k}\hat{w}_{i(j)}\hat{x}_j\right) - f(w_i^{\mathrm{T}}x)\right| \le C_L \left|\frac{1}{k}\sum_{j=1}^{k}\hat{w}_{i(j)}\hat{x}_j - w_i^{\mathrm{T}}x\right|.$$

Using Lemma 1 (see Section 14 for details and Appendix for the proof of the lemma) for vectors w_i and x we can bound the probability that the inner products before $(w_i^{\mathrm{T}}x)$ and after projection $(\frac{1}{k}\sum_{j=1}^{k}\hat{w}_{i(j)}\hat{x}_j)$ differ more than $\epsilon\|w_i\|\|x\|$, i.e.:

$$Pr\left\{\left|\frac{1}{k}\sum_{j=1}^{k}\frac{\hat{w}_{i(j)}}{\|w_i\|}\frac{\hat{x}_j}{\|x\|} - \left(\frac{w_i}{\|w_i\|}\right)^{\mathrm{T}}\frac{x}{\|x\|}\right| > \epsilon\right\}$$
$$= Pr\left\{\left|\frac{1}{k}\sum_{j=1}^{k}\hat{w}_{i(j)}\hat{x}_j - w_i^{\mathrm{T}}x\right| > \epsilon\|w_i\|\|x\|\right\} \le 2\exp -k\left(\frac{\epsilon^2}{4} - \frac{\epsilon^3}{6}\right),$$

where $\epsilon \in (0, 1)$.

Let $R = C_L \max_i |v_i|\|w_i\| \max_{x\in X}\|x\|$ and let $\epsilon = \frac{\varepsilon}{R}$. Thus, we obtain the bound

$$Pr\{|y(x) - y_S(x)| > \varepsilon\} \le 2M\exp\left[-k\left(\frac{\varepsilon^2}{4R^2} - \frac{\varepsilon^3}{6R^3}\right)\right]. \tag{4}$$

Selecting the admissible probability of error $\delta > 0$, we can obtain the lower bound on k. Since there are n data points in total, by the union bound inequality, it suffices if

$$nPr\{|y(x) - y_S(x)| > \varepsilon\} \le \delta.$$

Thus, we have

$$2nM\exp\left[-k\left(\frac{\varepsilon^2}{4R^2} - \frac{\varepsilon^3}{6R^3}\right)\right] \le \delta$$

and consecutively

$$\exp[\log(2nM)]\exp\left[-k\left(\frac{\varepsilon^2}{4R^2} - \frac{\varepsilon^3}{6R^3}\right)\right] \le \exp(\log\delta).$$

The last inequality holds if

$$\log(2nM) - \log(\delta) \le k\left(\frac{\varepsilon^2}{4R^2} - \frac{\varepsilon^3}{6R^3}\right),$$

which finishes the proof by setting $c_1 = \frac{1}{4R^2}$ and $c_2 = \frac{1}{6R^3}$. •

Note that R depends on maximal values of $\|x\|$, $\|w_i\|$ and $|v_i|$ — indicating that scaling of the data used for neural network learning is a very important task and may strongly influence the dimensionality reduction possibilities.

3. Linear random projections — a short review.

As a prerequisite for formulating and proving Lemma 1 we provide a review of results on random projections [12,1,13,14,2,15].

In random projections, we can estimate the original pairwise Euclidean distances directly using the corresponding Euclidean distances in a smaller dimension. Furthermore, the Johnson–Lindenstrauss lemma [4,16,7] provides the performance guarantee.

Let $u_i \in R^d$, $i = 1, \ldots, n$ be the original data. Let $S \in R^{k\times d}$ be a random matrix whose entries are i.i.d. samples of some random variable. The projected data are

$$v_i = Su_i \in R^k, \quad i = 1, \ldots, n.$$

Note that for $m, l \in \{1, \ldots, n\}$ we have

$$v_{mj} - v_{lj} = \sum_{i=1}^{d} s_{ji}(u_{mi} - u_{li}), \quad j = 1, 2, \ldots, k.$$

When $s_{ij} \sim N(0, 1)$ are i.i.d., then

$$v_{mj} - v_{lj} = \sum_{i=1}^{d} s_{ji}(u_{mi} - u_{li}) \sim N\left(0, \sum_{i=1}^{d}(u_{mi} - u_{li})^2\right),$$

and

$$X_j = \frac{v_{mj} - v_{lj}}{\left(\sum_{i=1}^{d}(u_{mi} - u_{li})^2\right)^{1/2}} \sim N(0, 1). \tag{5}$$

Denote by $\|.\|_{E(p)}$ the Euclidean distance in p dimensions. Then, we can estimate

$$\|u_m - u_l\|^2_{E(d)} = \sum_{i=1}^{d}(u_{mi} - u_{li})^2$$

from the sample squared distances as follows:

$$\hat{D}^2 = \frac{1}{k}\sum_{j=1}^{k}(v_{mj} - v_{lj})^2 = \|v_l - v_m\|^2_{E(k)}, \tag{6}$$

We omit displaying the dependence of $\hat{D}^2$ on m, l for the sake of brevity. Note that $\hat{D}^2$ is estimated using the elements of projected vectors. Below, we provide results, which show to what extent distances between projected points are close to distances of their counterparts in the original space.

It is easy to see that [7,2,13]:

$$E\{\hat{D}^2\} = D^2. \tag{7}$$

Thus, $\hat{D}^2$ is an unbiased estimator of "true" distance between points in the higher dimensional space, while its variance, presented below, decreases to zero as $k \to \infty$

$$var(\hat{D}^2) = \frac{2}{k}D^4. \tag{8}$$

Similarly,

$$E(\|v\|^2_{E(k)}) = \|u\|^2_{E(d)}, \quad var(\|v\|^2_{E(k)}) = \frac{2}{k}\|u\|^4_{E(d)}$$

for any $u \in R^d$ and $v = Su$ [2].

Furthermore, according to (5)

$$\frac{k\hat{D}}{D^2} = \sum_{i=j}^{k} X_j^2 \sim \chi_k^2,$$

where $\sum_{i=j}^{k} X_j^2 \sim \chi_k^2$ means that the sum has the chi-squared distribution with k degrees of freedom. Thus, using chi-squared tail Chernoff bounds (see [7] for details) we can obtain the bound on the probability when the relative error exceeds ε $(1 > \varepsilon > 0)$

$$Pr\left\{\frac{|\hat{D}^2 - D^2|}{D^2} \geq \varepsilon\right\} \leq 2\ \exp\left(-\frac{k}{4}\varepsilon^2 + \frac{k}{6}\varepsilon^3\right). \tag{9}$$

Obviously, the similar inequality holds for the vector's norms (a norm can be treated as a distance between given a vector and $[0, \ldots, 0]^{\mathrm{T}}$), i.e.,

$$Pr\left\{\frac{|\|v\|^2_{E(k)} - \|u\|^2_{E(d)}|}{\|u\|^2_{E(d)}} \geq \varepsilon\right\} \leq 2\ \exp\left(-\frac{k}{4}\varepsilon^2 + \frac{k}{6}\varepsilon^3\right). \tag{10}$$

In order to provide more explicit bounds, select the admissible probability of error $\delta > 0$. Since there are $n(n-1)/2 < n^2/2$ pairs in total among n data points, by the union bound inequality, it suffices if

$$\frac{n^2}{2}Pr\left\{|\hat{D}^2 - D^2| \geq \varepsilon D^2\right\} \leq \delta.$$

Using (9) we obtain

$$\frac{n^2}{2} 2 \exp\left(-\frac{k}{4}\varepsilon^2 + \frac{k}{6}\varepsilon^3\right) \le \delta$$

and consequently

$$k \ge \frac{2\log n - \log\delta}{\varepsilon^2/4 - \varepsilon^3/6} = c(n, \varepsilon, \delta)$$

provides the required upper bound for a probability of error.

It should be mentioned that one can also sample s_{ij} from other distributions with zero mean and unit variance (see [17, 14,6]).

The above inequalities are bounds for the probabilities of deviations between distances of pairs of points in the original space and in the projection space, which has a reduced dimension.

4. Random projections of inner products

Now we are interested in the properties of inner product operation after random projection S:

$$v_l^{\mathrm{T}} v_m = \sum_{i=1}^{k} v_{li} v_{mi}$$

in comparison to the original value $u_l^{\mathrm{T}} u_m$.

It is relatively easy to see [5,6] that

$$E(v_{li} v_{mi}) = u_l^{\mathrm{T}} u_m, \quad i = 1, \ldots, k.$$

Let $I(u_l, u_m) = u_l^{\mathrm{T}} u_m$
and

$$\hat{I}(u_l, u_m) = \frac{1}{k} \sum_{i=1}^{k} v_{li} v_{mi}.$$

Thus,

$$E(\hat{I}(u_l, u_m)) = \frac{1}{k} \sum_{i=1}^{k} E(v_{li} v_{mi}) = u_l^{\mathrm{T}} u_m. \tag{11}$$

Furthermore [5,6],

$$var(v_{li} v_{mi}) = (u_l^{\mathrm{T}} u_m)^2 + \|u_l\|^2 \|u_m\|^2, \quad i = 1, \ldots, k$$

and

$$var(\hat{I}(u_l, u_m)) = \frac{1}{k}((u_l^{\mathrm{T}} u_m)^2 + \|u_l\|^2 \|u_m\|^2). \tag{12}$$

Using the Chebyshev inequality we can obtain

$$\begin{aligned}
&Pr\left\{\frac{|\hat{I}(u_l, u_m) - I(u_l, u_m)|}{\|u_l\|\|u_m\|} \ge \epsilon\right\} \\
&= Pr\left(\frac{\left|\frac{1}{k} v_l^{\mathrm{T}} v_m - E\left(\frac{1}{k} v_l^{\mathrm{T}} v_m\right)\right|}{\|u_l\|\|u_m\|} \ge \epsilon\right) \\
&= Pr\left(\frac{\left|\sum_{i=1}^{k} v_{li} v_{mi} - (u_l^{\mathrm{T}} u_m)\right|}{\|u_l\|\|u_m\|} \ge k\epsilon\right) \le \frac{var\left(\sum_{i=1}^{k} v_{li} v_{mi}\right)}{k^2 \epsilon^2 \|u_l\|^2 \|u_m\|^2} \\
&= \frac{k^2 \frac{1}{k}((u_l^{\mathrm{T}} u_m)^2 + \|u_l\|^2 \|u_m\|^2)}{k^2 \epsilon^2 \|u_l\|^2 \|u_m\|^2} \le \frac{2}{k\epsilon^2}
\end{aligned} \tag{13}$$

since $u_l^{\mathrm{T}} u_m \le \|u_l\|\|u_m\|$.

Denote by r the inner product of normed vectors $x_l = u_l/\|u_l\|$ and $x_m = u_m/\|u_m\|$, i.e.,

$$r = I(x_l, x_m) = \frac{u_l^{\mathrm{T}} u_m}{\|u_l\|\|u_m\|}.$$

Thus, we obtain a simpler form of inequality (13)

$$Pr\left\{|\hat{I}(x_l, x_m) - r| \geq \epsilon\right\} \leq \frac{2}{k\epsilon^2} \tag{14}$$

Note, that (14) differs from inequality (9) (formulated for distances). We cannot evaluate $|\hat{I}(x_l, x_m) - r|/r$, because $r \in [-1, 1]$ and the factor $1/r^2$, which, in such a case, would appear in the right-hand side of inequality (14), can be very large for r close to zero.

Furthermore, selecting the admissible probability of error $\delta > 0$, and using the Bonferroni union bound inequality

$$\frac{n^2}{2} Pr\left\{|\hat{I}(x_l, x_m) - r| \geq \epsilon\right\} \leq \delta.$$

we obtain the inequality

$$\frac{n^2}{k\epsilon^2} \leq \delta.$$

Thus, selecting $k \geq \frac{n^2}{\delta\epsilon^2}$, we get a crude lower bound for k.

The other possibility is to use the Chernoff bounds instead of the Chebyshev inequality for bounding the probability

$$P = Pr\left\{|\hat{I}(x_l, x_m) - r| \geq \epsilon\right\}.$$

Using the Chernoff bounds we are able to show the following lemma. The proof of the lemma is given in details in the Appendix.

Lemma 1. *If $\hat{x_m}$ and $\hat{x_l}$ are random normal linear projection of vectors $x_m \in R^d$ and $x_l \in R^d$ onto R^k then*

$$Pr\left\{|\hat{I}(x_l, x_m) - r| > \epsilon\right\}$$
$$= Pr\left\{\left|\frac{1}{k}\sum_{j=1}^{k} \frac{\hat{x_{m(j)}}}{\|x_m\|}\frac{\hat{x_{l(j)}}}{\|x_l\|} - \left(\frac{x_m}{\|x_m\|}\right)^{\mathrm{T}} \frac{x_l}{\|x_l\|}\right| > \varepsilon\right\} \leq 2\exp\left[-k\left(\frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}\right)\right],$$

where $\varepsilon \in (0, 1)$.

5. Conclusions

We have shown that normal random projections preserve approximately, not only the squared Euclidean distance between pairs of projected points, but also their inner products. The concentration of inner product property formulated in Lemma 1 allows us to justify introducing the dimensionality reduction layer as a first layer in the feed-forward neural network with the sigmoidal activation functions.

It seems that a similar approach could also be employed for radial basis function networks, but the large values of the Lipschitz constant may lead to a very large lower bound for k.

Acknowledgments

This work was supported by a grant of the Ministry of Science and Higher Education for 2006-2009.

Appendix. Proof of Lemma 1

We use the Chernoff bounds for bounding the probability

$$P = Pr\left\{|\hat{I}(x_l, x_m) - r| > \epsilon\right\}.$$

Let $z_l = v_l/\|u_l\| = Sx_l$. The probability $P = P_+ + P_-$, where

$$P_+ = Pr\left\{\sum_{i=1}^{k} z_{li}z_{mi} - rk > k\epsilon\right\},$$

and

$$P_- = Pr\left\{\sum_{i=1}^{k} z_{li}z_{mi} - rk < -k\epsilon\right\}.$$

Using the Chernoff bound we obtain for any s that

$$\begin{aligned} P_+ &= Pr\left\{\exp\left(s\sum_{i=1}^{k} z_{li}z_{mi} - ks(r+\epsilon)\right] > 1\right\} \\ &\le \exp(-ks(r+\epsilon))E\left[\exp\left(s\sum_{i=1}^{k} z_{li}z_{mi}\right)\right] = \exp[-ks(r+\epsilon)]\Pi_{i=1}^{k}E[\exp(sz_{li}z_{mi})] \end{aligned}$$

and

$$\begin{aligned} P_- &= Pr\left\{\exp\left[-s\sum_{i=1}^{k} z_{li}z_{mi} + ks(r-\epsilon)\right] > 1\right\} \\ &\le \exp[ks(r-\epsilon)]E\left[\exp\left(-s\sum_{i=1}^{k} z_{li}z_{mi}\right)\right] = \exp[ks(r-\epsilon)]\Pi_{i=1}^{k}E[\exp(-sz_{li}z_{mi})]. \end{aligned}$$

Random variables z_{li} i z_{mi} are correlated and for each i they have the same probability distribution which is a bivariate normal with the mean $[0, 0]^{\mathrm{T}}$ and the correlation matrix:

$$\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}.$$

Note, that for the sake of simplicity of the notation, the same symbol stands for random variables and for their values.

Define $y_{1i} = \sqrt{2}/2z_{li} + \sqrt{2}/2z_{mi}$ and $y_{2i} = -\sqrt{2}/2z_{li} + \sqrt{2}/2z_{mi}$.

It is easy to see that y_{1i} and y_{2i}, $(i = 1, \ldots, k)$ are normal random variables, which are mutually independent. Thus, $y_{1i} \sim N[0, 1+r]$ and $y_{2i} \sim N[0, 1-r]$.

Hence, the moment generation function

$$E[\exp(s\, z_{li}z_{mi})]$$

is equal to

$$E[\exp(s/2(y_{1i}^2 - y_{2i}^2)] = E[\exp(s/2(y_{1i}^2)]E[\exp(-s/2(y_{2i}^2)],$$

where $y_{1i}^2/(1+r),\ y_{2i}^2/(1-r) \sim \chi_1^2$.

It is known that $E(\exp(sX)) = \sqrt{1-2s}$, if $X \sim \chi_1^2$, so

$$E[\exp(s\, z_{li}z_{mi})] = ((1-s(1+r))(1+s(1-r))^{-1/2}, \quad r \in [-1, 1].$$

Thus,

$$P_+ \le [(1-s(1+r))(1+s(1-r)]^{-k/2}\exp(-ks(r+\epsilon)),$$

where $\frac{-1}{1-r} \le s \le \frac{1}{1+r}$ and

$$P_- \le \exp(ks(r-\epsilon))[(1+s(1+r))(1-s(1-r)]^{-k/2},$$

where $\frac{-1}{1+r} \le s \le \frac{1}{1-r}$.

We start with bounding P_+. Assume that $s = \frac{\epsilon}{2(1+\epsilon r)}$. Note, that $-2/(1+r) \le \epsilon \le 2/(1-r)$ hold for every $\epsilon \in (0, 1)$ and $r \in [-1, 1]$.

$$\begin{aligned} P_+ &\le \exp\left[\frac{k}{2}\left(-2\frac{\epsilon}{2(1+\epsilon r)}(r+\epsilon) - \log\left[1 - \frac{\epsilon}{2(1+\epsilon r)}(1+r)\right] - \log\left[1 + \frac{\epsilon}{2(1+\epsilon r)}(1-r)\right]\right)\right] \\ &= \exp\left[\frac{k}{2}\left(-\frac{\epsilon}{1+r\epsilon}(r+\epsilon) - \log\left[1 - \frac{\epsilon(1+r)}{2(1+\epsilon r)}\right] - \log\left[1 + \frac{\epsilon(1-r)}{2(1+\epsilon r)}\right]\right)\right]. \end{aligned}$$

We have to show that function

$$f_+(\epsilon, r) = -\frac{\epsilon}{1+r\epsilon}(r+\epsilon) - \log\left[1 - \frac{\epsilon(1+r)}{2(1+\epsilon r)}\right] - \log\left[1 + \frac{\epsilon(1-r)}{2(1+\epsilon r)}\right]$$

is negative for $r \in [-1, 1]$ and $\epsilon \in (0, 1)$. Note, that $f_+(\epsilon, r)$ is continuous for $r \in [-1, 1]$ and $\epsilon \in (0, 1)$.

If $r = 1$, then

$$\begin{aligned} f_+(\epsilon, 1) &= -\epsilon - \log\left[1 - \frac{\epsilon(1+r)}{2(1+\epsilon r)}\right] - \log\left[1 + \frac{\epsilon(1-r)}{2(1+\epsilon r)}\right] \\ &= -\epsilon - \log\left[1 - \frac{\epsilon}{1+\epsilon}\right] - \log 1 = -\epsilon - \log\frac{1}{1+\epsilon} \\ &= -\epsilon + \log(1+\epsilon) \le -\epsilon + \epsilon - \epsilon^2/2 + \epsilon^3/3 \\ &= -\epsilon^2/2 + \epsilon^3/3 < 0, \end{aligned}$$

provided that $\epsilon < 1$.

Similarly, if $r = -1$, then

$$f_+(\epsilon, -1) = \epsilon - \log 1 - \log\left[1 + \frac{\epsilon}{(1-\epsilon)}\right] = \epsilon + \log(1-\epsilon) \le \epsilon - \epsilon - \epsilon^2/2 < 0,$$

provided that $\epsilon < 1$.

Furthermore, the derivative of $f_+(\epsilon, r)$ with respect to r is the following

$$\begin{aligned} f_+^{'}(\epsilon, r) &= \frac{\epsilon^2\left(2r + \epsilon\left(\left(r^2-1\right)\epsilon^2 + 2r\epsilon + r^2 + 3\right)\right)}{(\epsilon(r-1)+2)(\epsilon r+1)^2(r\epsilon+\epsilon+2)} \\ &= \frac{\epsilon^2[r^2(^3+\epsilon) + 2r(1+\epsilon^2) + 3\epsilon - \epsilon^3]}{(\epsilon(r-1)+2)(\epsilon r+1)^2(\epsilon r+\epsilon+2)} = \frac{\epsilon^2 L_+(\epsilon, r)}{M_+(\epsilon, r)}, \end{aligned}$$

where $L_+(\epsilon, r)$ and $M_+(\epsilon, r)$ denote the corresponding part of numerator and the denominator, respectively. It is easy to see that $M_+(\epsilon, r) > 0$ for $r \in [-1, 1]$ and $\epsilon \in (0, 1)$.

Furthermore, $f_+^{'}(\epsilon, r) = 0$ for

$$r_1(\epsilon) = \frac{-\epsilon^2 - \sqrt{\epsilon^6 - \epsilon^4 - \epsilon^2 + 1} - 1}{\epsilon^3 + \epsilon} < -1, \quad \epsilon \in (0, 1)$$

and

$$r_2(\epsilon) = \frac{-\epsilon^2 + \sqrt{\epsilon^6 - \epsilon^4 - \epsilon^2 + 1} - 1}{\epsilon^3 + \epsilon} < 0,$$

where $r_2(\epsilon) > -1$, if $\epsilon < 1$ and $r_2(\epsilon) -> 0$ if $\epsilon -> 0$.

To the end, observe that $L_+(\epsilon, r) < 0$ (and consequently $f_+^{'}(\epsilon, r) < 0$), if $r \in [-1, r_2(\epsilon)]$, $\epsilon \in (0, 1)$. Similarly, $L_+(\epsilon, r) > 0$, if $r \in [r_2(\epsilon), 1]$, $\epsilon \in (0, 1)$. Summarizing, we have shown that $f_+(\epsilon, r) < 0$ if $r \in [-1, 1]$, $\epsilon \in (0, 1)$.

As a consequence, we obtain that

$$P_+ \le \exp\left[-\frac{k}{4}\epsilon^2 + \frac{k}{6}\epsilon^3\right].$$

For the second Chernoff bound let us assume that $s = \epsilon/(2-2\epsilon r)$. Consecutively, we obtain $-2/(1+r) \le \epsilon \le 2/(1-r)$, which is true for any $\epsilon \in (0, 1)$ and $r \in [-1, 1]$.

Thus,

$$\begin{aligned} P_- &\le \exp\left[\frac{k}{2}\left(\frac{\epsilon}{1-\epsilon r}(r-\epsilon) - \log\left[1 + \frac{\epsilon}{2(1-\epsilon r)}(1+r)\right] - \log\left[1 - \frac{\epsilon}{2(1-\epsilon r)}(1-r)\right]\right)\right] \\ &= \exp\left[\frac{k}{2}\left(\frac{\epsilon}{1-r\epsilon}(r-\epsilon) - \log\left[1 + \frac{\epsilon(1+r)}{2(1-\epsilon r)}\right] - \log\left[1 - \frac{\epsilon(1-r)}{2(1-\epsilon r)}\right]\right)\right]. \end{aligned}$$

We have to show that function

$$f_-(\epsilon, r) = \frac{\epsilon}{1-r\epsilon}(r-\epsilon) - \log\left[1 + \frac{\epsilon(1+r)}{2(1-\epsilon r)}\right] - \log\left[1 - \frac{\epsilon(1-r)}{2(1-\epsilon r)}\right]$$

is negative for $r \in [-1, 1]$ and $\epsilon \in (0, 1)$. Note, that $f_-(\epsilon, r)$ is a continuous function for $r \in [-1, 1]$ and $\epsilon \in (0, 1)$.

If $r = 1$, then

$$\begin{aligned} f_-(\epsilon, 1) &= \epsilon - \log\left[1 + \frac{\epsilon}{1-\epsilon}\right] - \log 1 = \epsilon - \log\frac{1}{1-\epsilon} \\ &= \epsilon + \log(1-\epsilon) \le \epsilon - \epsilon - \epsilon^2/2 \\ &= -\epsilon^2/2 < 0, \end{aligned}$$

provided that $\epsilon < 1$.

Similarly, if $r = -1$, then

$$f_-(\epsilon, -1) = -\epsilon - \log 1 - \log\left[1 - \frac{\epsilon}{(1+\epsilon)}\right] = -\epsilon + \log(1-\epsilon)$$
$$\leq -\epsilon + \epsilon - \epsilon^2/2 + \epsilon^3/3 = -\epsilon^2/2 + \epsilon^3/3 < 0$$

provided that $\epsilon \in (0, 1)$.

Furthermore, the derivative of $f_-(\epsilon, r)$ with respect to r is given by

$$f_-^{'}(\epsilon, r) = \frac{\epsilon^2\left(2r - \epsilon\left(\left(r^2-1\right)\epsilon^2 - 2r\epsilon + r^2 + 3\right)\right)}{(\epsilon(r-1)-2)(\epsilon r - 1)^2(r\epsilon + \epsilon - 2)}$$
$$= \frac{\epsilon^2[-r^2(\epsilon^3+\epsilon) + 2r(1+\epsilon^2) - 3\epsilon + \epsilon^3]}{(\epsilon r + \epsilon - 2)(\epsilon r - 1)^2(\epsilon r - \epsilon - 2)} = \frac{\epsilon^2 L_-(\epsilon, r)}{M_-(\epsilon, r)},$$

where $L_-(\epsilon, r)$ and $M_-(\epsilon, r)$ denote the corresponding part of numerator and the denominator, respectively. It is easy to see that $M_-(\epsilon, r) > 0$ for $r \in [-1, 1]$ and $\epsilon \in (0, 1)$.

Additionally, $f_-^{'}(\epsilon, r) = 0$ for

$$r_1(\epsilon) = \frac{\epsilon^2 - \sqrt{\epsilon^6 - \epsilon^4 - \epsilon^2 + 1} + 1}{\epsilon^3 + \epsilon} > 0, \quad \epsilon \in (0, 1)$$

and $r_1(\epsilon) > 0$. Note, that $r_2(\epsilon) > 1$.

To the end, $L_-(\epsilon, r) < 0$ (and $f_-^{'}(\epsilon, r) < 0$), if $r \in [-1, r_1(\epsilon)]$ and $\epsilon \in (0, 1)$. Similarly, $L_-(\epsilon, r) > 0$, if $r \in [r_1(\epsilon), 1]$, $\epsilon \in (0, 1)$. Summarizing, we have shown, that

$$f_-(\epsilon, r) < 0 \quad \text{if } r \in [-1, 1], \quad \epsilon \in (0, 1).$$

As a consequence, we obtain that also

$$P_- \leq \exp\left[-\frac{k}{4}\epsilon^2 + \frac{k}{6}\epsilon^3\right].$$

Adding bounds for P_+ and P_- we arrive at

$$P \leq 2\exp\left[-\frac{k}{4}\epsilon^2 + \frac{k}{6}\epsilon^3,\right]$$

which finishes the proof. •

References

[1] R. Arriaga, S. Vempala, An algorithmic theory of learning: Robust concepts and random projection, in: Proc. of FOCS 1999, New York, 1999, pp. 616–623.
[2] S. Vempala, The Random Projection Method, American Mathematical Society, Providence, RI, 2004.
[3] M. Anthony, P.L. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, Cambridge, 1999.
[4] W.B. Johnson, J. Lindenstrauss, Extensions of Lipshitz mapping into Hilbert space, Contemporary Mathematics 26 (1984) 189–206.
[5] 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 (2006) 92–106.
[6] P. Li, T.J. Hastie, K.W. Church, Improving Random Projections Using Marginal Information, in: LNCS, vol. 4005/2006, 2006, pp. 635–649.
[7] S. Dasgupta, A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss, Random Structures and Algorithms 22 (1) (2003) 60–65.
[8] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.
[9] S. Haykin, Neural Networks. A Comprehensive Foundation, 2nd ed., Prentice-Hall, New Jersey, 1999.
[10] G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals, and Systems 2 (1989) 303–314.
[11] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (1989) 359–366.
[12] P. Indyk, R. Motwani, Approximate nearest neighbors: Towards removing the curse of dimensionality, in: Proc. of STOC 1998, Dallas, TX, 1998, pp. 604–613.
[13] P. Li, T.J. Hastie, K.W. Church, Nonlinear estimators and tail bounds for dimension reduction in l1 using Cauchy random projections, The Journal of Machine Learning Research 8 (10) (2007) 2497–2532.
[14] D. Achlioptas, Database friendly random projections, in: Proc. Principles of Database Systems, PODS, 2001, pp. 274–281.
[15] N. Ailon, B. Chazelle, Approximate nearest neighbors and the fast Johnson—Lindenstrauss transform, in: Proc. of STOC 2006, Seattle, WA, 2006, pp. 557–563.
[16] P. Frankl, H. Maehara, The Johnson—Lindenstrauss lemma and the sphericity of some graphs, Journal of Combinatorial Theory A 44 (3) (1987) 355–362.
[17] P. Indyk, A. Naor, Nearest neighbor preserving embeddings, ACM Transactions on Algorithms(TALG)archive 3 (2007) Article No. 31.