ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 12, 2008

Free-lunch learning and directional distributions in artificial neural networks

P. E. Jupp and J. V. Stone

Abstract. Free-lunch learning (FLL) is a phenomenon in which relearning partially-forgotten mental associations induces recovery of other associations. FLL occurs also in artificial neural networks (ANN's). Two models for forgetting in ANN's are presented which involve uniform distributions on spheres and Grassmann manifolds. It is shown that these models differ markedly in their amounts of FLL.

1. Introduction

Most people who have learned a foreign language have experienced the following sequence of events:

(i) one learns two sets of vocabulary, A_1 and A_2, in the foreign language;
(ii) one forgets (possibly partially) $A_1 \cup A_2$;
(iii) one relearns *only* A_2 — and then finds that knowledge of A_1 is (to some extent) recovered.

This phenomenon, in which restoration of A_1 comes "free" with the relearning of A_2 is called *free-lunch learning* (FLL). (The popular aphorism "There's no such thing as a free lunch" seems not to apply in learning theory.)

Free-lunch learning occurs in contexts other than language. Stone *et al.* (2001) demonstrated FLL using a task in which participants learned the positions of letters on a non-standard computer keyboard. After a period of forgetting, participants relearned a proportion of these positions. It was found that this relearning induced recovery of the positions that had not been relearned. Stone (2007) has shown that FLL accelerates evolution of adaptive behaviours.

Stone and Jupp (2007, 2008) showed that FLL can occur also in artificial neural networks (ANN's), and investigated its behaviour under two

Received October 27, 2007.
2000 *Mathematics Subject Classification.* 92B20, 62H11.
Key words and phrases. Grassmann manifold, isotropy, sphere, uniform distribution.

contrasting models of the forgetting process. The aims of this paper are (i) to draw the attention of statisticians to the role of directional statistics in artificial neural networks, (ii) to extend the results of Stone and Jupp (2007, 2008) on FLL by weakening the distributional assumptions made there.

Section 2 describes FLL in ANN's and introduces a measure of FLL. Section 3 considers a model of forgetting in ANN's in which forgetting is due to perturbation of the weight vector. It is shown that for large networks, under appropriate isotropy conditions, FLL is very probable. Section 4 considers a model in which forgetting is due to the weight vector "fading" towards the origin. Under appropriate isotropy conditions on large networks, the probability of FLL is almost zero.

2. Free-lunch learning in artificial neural networks

A very simple mathematical model for the activity of a human brain is given by an elementary artificial neural network. This is a directed graph with nodes (vertices) corresponding to the neurons of the brain and directed edges corresponding to the synapses. The state of any node is is a real number which models the firing rate of the corresponding neuron. (Background on ANN's can be found in e.g. Bishop, 1995, Müller and Reinhardt, 1990, and Titterington, 1999.) It is enough here to consider *linear* ANN's with n input nodes and a single output node; the ANN transforms an input vector $\mathbf{x}$ in $\mathbb{R}^n$ into the output $\mathbf{w}^T\mathbf{x}$, where the *weight vector* $\mathbf{w}$ in $\mathbb{R}^n$ represents the state of the ANN. (This linear one-layer model is almostly certainly too simple to represent adequately the complexity of a human brain. Nevertheless, it will be shown in Section 3 that even this simple model can display FLL with high probability.)

For input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_c$ and desired outputs $d_1, \ldots, d_c$ the squared error is defined as

$$\sum_{i=1}^{c} \left(\mathbf{w}^T\mathbf{x}_i - d_i\right)^2 = \|\mathbf{X}\mathbf{w} - \mathbf{d}\|^2,$$

where

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_c)^T, \qquad \mathbf{d} = (d_1, \ldots, d_c)^T.$$

Teaching the ANN to associate inputs $\mathbf{x}_1, \ldots, \mathbf{x}_c$ with respective outputs $d_1, \ldots, d_c$ puts the weight vector $\mathbf{w}$ into the affine subspace $\{\mathbf{w} : \mathbf{X}\mathbf{w} = \mathbf{d}\}$ of $\mathbb{R}^n$.

The description of FLL in language-learning that was given in the Introduction can be translated into the following version for ANN's:

(i) the ANN learns two sets of associations, A_1 and A_2, that associate to inputs $\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}$ the outputs $d_{11}, \ldots, d_{1n_1}$ and associate to inputs $\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2}$ the outputs $d_{21}, \ldots, d_{2n_2}$, respectively. Denote the weight vector after learning A_1 and A_2 by $\mathbf{w}_0$;

(ii) the ANN forgets (possibly partially) $A_1 \cup A_2$. This moves the weight vector to $\mathbf{w}_1$, say. The squared error on A_1 is then $\|\mathbf{X}_1\mathbf{w}_1 - \mathbf{d}_1\|^2$;
(iii) the ANN relearns *only* A_2. This moves the weight vector to $\mathbf{w}_2$, the orthogonal projection of $\mathbf{w}_1$ onto the affine subspace $\{\mathbf{w} : \mathbf{X}_1\mathbf{w} = \mathbf{d}_1\}$. The squared error on A_1 is then $\|\mathbf{X}_1\mathbf{w}_2 - \mathbf{d}_1\|^2$.

A useful measure of the amount of FLL is

$$\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) = \|\mathbf{X}_1\mathbf{w}_1 - \mathbf{d}_1\|^2 - \|\mathbf{X}_1\mathbf{w}_2 - \mathbf{d}_1\|^2.$$

Free-lunch learning has occurred if $\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0$.

For $i = 1, 2$ the $n_i \times n$ matrix $\mathbf{X}_i$ can be written as

$$\mathbf{X}_i = \mathbf{T}_i\mathbf{Z}_i$$

for $n_i \times n_i$ and $n_i \times n$ matrices $\mathbf{T}_i$ and $\mathbf{Z}_i$ such that $\mathbf{T}_i$ is lower triangular with positive diagonal elements and $\mathbf{Z}_i\mathbf{Z}_i^T = \mathbf{I}_{n_i}$. If $\mathbf{X}_i$ has rank n_i then $\mathbf{T}_i$ and $\mathbf{Z}_i$ are unique. The matrix $\mathbf{Z}_i^T\mathbf{Z}_i$ represents the operator which projects onto the image of $\mathbf{X}_i^T\mathbf{X}_i$. Thus, if $\mathbf{X}_i$ has rank n_i, $\mathbf{Z}_i^T\mathbf{Z}_i$ is an element of the Grassmann manifold $G_{n_i}(\mathbb{R}^n)$ of (orthogonal projections onto) n_i-dimensional subspaces of $\mathbb{R}^n$.

Algebraic manipulation shows that if $(\mathbf{X}_1, \mathbf{d}_1)$ and $(\mathbf{X}_2, \mathbf{d}_2)$ are consistent (i.e. there is a $\mathbf{w}_0$ satisfying $\mathbf{X}_i\mathbf{w}_0 = \mathbf{d}_i$ for $i = 1, 2$) then

$$\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) = \mathbf{v}^T\mathbf{Z}_2^T\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\left(2\mathbf{I}_n - \mathbf{Z}_2^T\mathbf{Z}_2\right)\mathbf{v}, \tag{2.1}$$

where $\mathbf{v}$ is the "forgetting vector"

$$\mathbf{v} = \mathbf{w}_1 - \mathbf{w}_0.$$

From now on $(\mathbf{X}_1, \mathbf{d}_1)$, $(\mathbf{X}_2, \mathbf{d}_2)$ and $\mathbf{v}$ will be random and the assumptions

(A1) $n_1 + n_2 \leq n$,
(A2) $(\mathbf{X}_1, \mathbf{d}_1)$ and $(\mathbf{X}_2, \mathbf{d}_2)$ have continuous distributions (i.e. have densities with respect to Lebesgue measure)

will be made. These assumptions ensure that, with probability 1, $\mathbf{X}_i$ has rank n_i for $i = 1, 2$ and (2.1) holds.

The results in Sections 3 and 4 are based on the following concepts of isotropy:

(i) $\mathbf{v}$ has an *isotropic* distribution on $\mathbb{R}^n$ if $\mathbf{v}$ has the same distribution as $\mathbf{U}\mathbf{v}$ for all orthogonal $n \times n$ matrices $\mathbf{U}$, i.e. the corresponding unit vector $\|\mathbf{v}\|^{-1}\mathbf{v}$ has the uniform distribution on S^{n-1} (see §9.3.1 of Mardia and Jupp, 2000);
(ii) $\mathbf{Z}_i^T\mathbf{Z}_i$ is *uniformly* distributed on the Grassmann manifold $G_{n_i}(\mathbb{R}^n)$ if $\mathbf{Z}_i^T\mathbf{Z}_i$ has the same distribution as $\mathbf{W}^T\mathbf{Z}_i^T\mathbf{Z}_i\mathbf{W}$ for all orthogonal $n_i \times n_i$ matrices $\mathbf{W}$ (see §13.3 of Mardia and Jupp, 2000).

3. Synaptic drift

A plausible model for the process of forgetting is the simple *synaptic drift model* introduced in Stone and Jupp (2007), in which the forgetting vector $\mathbf{v}$ is isotropic and independent of the weight vector $\mathbf{w}_0$ achieved by learning A_1 and A_2. It is useful here to extend the definition by using the term *synaptic drift model* to mean a model in which either $\mathbf{v}|\,(\mathbf{X}_1, \mathbf{Z}_2)$ is isotropic or $\mathbf{X}_1|\,(\mathbf{Z}_2, \mathbf{v})$ has identically-distributed isotropic rows.

The main properties of FLL in the synaptic drift model are given in the following theorems. The proof of Theorem 1 is analogous to that in §A.5.1 of Appendix A of Stone and Jupp (2007). Proofs of Theorems 2 and 3 are given in the Appendix.

Theorem 1. *If* $\mathbf{v}|\,(\mathbf{X}_1, \mathbf{d}_1, \mathbf{Z}_2, \mathbf{d}_2)$ *is isotropic then*

$$\text{median}\,(\,\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1)|\,\mathbf{X}_1, \mathbf{d}_1, \mathbf{X}_2, \mathbf{d}_2) > 0.$$

Theorem 2. *The following is true:*

(i) *If* $\mathbf{v}|\,(\mathbf{X}_1, \mathbf{Z}_2)$ *is isotropic then*

$$\mathrm{E}\left[\,\delta\left(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1\right)|\,\mathbf{X}_1, \mathbf{Z}_2\right] = \frac{\mathrm{E}\left[\,\|\mathbf{v}\|^2\big|\,\mathbf{X}_1, \mathbf{Z}_2\right]}{n}\mathrm{tr}\left(\left(\mathbf{Z}_2\mathbf{X}_1^T\right)^T\left(\mathbf{Z}_2\mathbf{X}_1^T\right)\right),$$

and so

$$\mathrm{E}\left[\delta\left(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1\right)\right] \geq 0.$$

(ii) *If* $\mathbf{X}_1|\,(\mathbf{Z}_2, \mathbf{v})$ *has identically-distributed isotropic rows then*

$$\mathrm{E}\left[\,\delta\left(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1\right)|\,\mathbf{Z}_2, \mathbf{v}\right] = \frac{n_1\mathrm{E}\left[\,\|\mathbf{x}\|^2\big|\,\mathbf{Z}_2, \mathbf{v}\right]}{n}\|\mathbf{Z}_2\mathbf{v}\|^2,$$

where $\mathbf{x}$ *is any column of* $\mathbf{X}_1^T$*, and so*

$$\mathrm{E}\left[\delta\left(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1\right)\right] \geq 0.$$

Theorem 3. *One has:*

(i) *If* $\mathbf{v}|\,(\mathbf{X}_1, \mathbf{Z}_2)$ *is isotropic then*

$$P\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) < 0\right) \leq \frac{4}{n_2}.$$

(ii) *If* $\mathbf{X}_1$*,* $\mathbf{Z}_2$ *and* $\mathbf{v}$ *are independent,* $\mathbf{X}_1$ *has independent identically-distributed isotropic rows, and* ${\mathbf{Z}_2}^T\mathbf{Z}_2$ *is uniform then*

$$P\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \leq 0\right) \leq \frac{a_0(n, n_1, n_2) + a_1(n, n_2)\gamma(n)}{n_1 n_2 (n+2)^2},$$

where

$$\begin{aligned} a_0(n, n_1, n_2) &= 2(2n^3 - n^2 n_2 + 3n^2 - 2nn_2 - 2n_2) \\ a_1(n, n_2) &= n^2(4n - n_2 + 6) \\ \gamma(n) &= \frac{\operatorname{var}\left(\|\mathbf{x}\|^2\right)}{E\left[\|\mathbf{x}\|^2\right]^2}, \end{aligned}$$

with the n-vector $\mathbf{x}$ denoting any column of $\mathbf{X}_1^T$. If, further, n_1/n, n_2/n and $\gamma(n)/n$ are bounded away from zero as $n \to \infty$ then

$$P\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0\right) \to 1, \qquad n \to \infty.$$

Thus, under synaptic drift, if there are many nodes then the probability of free-lunch learning is very high.

4. Synaptic fading

A reasonable alternative to the synaptic drift model is the *synaptic fading model* introduced in Stone and Jupp (2008), in which the weight vector $\mathbf{w}_1$ after forgetting is given by

$$\mathbf{w}_1 = r\mathbf{w}_0, \tag{4.1}$$

for some (random) scalar r. Thus the "forgetting vector" $\mathbf{v}$ is given by

$$\mathbf{v} = (1 - r)\mathbf{w}_0.$$

The interpretation of (4.1) is that forgetting consists of shrinking the weight vector $\mathbf{w}_0$ by a factor r towards the "dead state" $\mathbf{0}$. Algebraic manipulation shows that

$$\begin{aligned} \delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) = (1 - r)^2 \Big\{ & 2\left(\mathbf{T}_2{}^{-1}\mathbf{d}_2\right)^T \mathbf{Z}_2 \mathbf{X}_1^T \mathbf{d}_1 \\ & - \left(\mathbf{T}_2{}^{-1}\mathbf{d}_2\right)^T \mathbf{Z}_2 \mathbf{X}_1{}^T \mathbf{X}_1 \mathbf{Z}_2^T \mathbf{T}_2{}^{-1}\mathbf{d}_2 \Big\}. \end{aligned} \tag{4.2}$$

The main properties of FLL in the synaptic fading model are given in the following theorems. Proofs of Theorems 4 and 5 are given in the Appendix.

Theorem 4. *If $\mathbf{d}_1 | (\mathbf{X}_1, \mathbf{X}_2, \mathbf{d}_2, r)$ and $-\mathbf{d}_1 | (\mathbf{X}_1, \mathbf{X}_2, \mathbf{d}_2, r)$ have the same distribution then*

$$\mathrm{E}\left[\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \,|\mathbf{X}_1, \mathbf{X}_2, \mathbf{d}_2, r\right] = -(1 - r)^2 \|\mathbf{X}_1 \mathbf{Z}_2^T \mathbf{T}_2^{-1} \mathbf{d}_2\|^2, \tag{4.3}$$

and so

$$\mathrm{E}\left[\delta\left(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1\right)\right] \leq 0.$$

Theorem 5. *If*

(a) $\mathbf{X}_1$, $\mathbf{d}_1$, $\mathbf{X}_2$, $\mathbf{d}_2$ *and r are independent,*
(b) $\mathbf{d}_1$ *and* $\mathbf{d}_2$ *are isotropic,*
(c) $\mathbf{X}_1$ *has independent identically-distributed isotropic rows and their density is bounded,*

(d) $\mathbf{X}_2$ *has identically-distributed rows*

then

$$P\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0\right) \leq \frac{12K\mathrm{E}\left[\|\mathbf{d}_1\|^2\right]\mathrm{E}\left[\|\tilde{\mathbf{x}}\|^2\right]\mathrm{E}\left[\|\mathbf{d}_2\|^{-2}\right]}{n_1(n_1-3)\mathrm{E}\left[\|\mathbf{x}\|^2\right]}, \qquad (4.4)$$

where $\mathbf{x}$ *and* $\tilde{\mathbf{x}}$ *denote arbitrary columns of* $\mathbf{X}_1^T$ *and* $\mathbf{X}_2^T$*, respectively, and*

$$K = \mathrm{E}\left[\frac{1}{t_1^2 + t_2^2 + t_3^2}\right]$$

with

$$t_i = \frac{1}{\sqrt{\mathrm{E}\left[\|\mathbf{x}_i\|^2\right]}}\mathbf{x}_i{}^T\mathbf{u} \quad \textit{for } i = 1, 2, 3,$$

$\mathbf{u}$ *being any unit vector in* $\mathbb{R}^n$*.*

Corollary. *If the conditions of Theorem 5 hold and*

$$\mathbf{d}_1 \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n_1}), \quad \mathbf{d}_2 \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n_2}),$$

$$\mathbf{x}_i, \tilde{\mathbf{x}}_j \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n) \quad \textit{for } i = 1, \ldots, n_1,\ j = 1, \ldots, n_2$$

then

$$P\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0\right) \leq \frac{12n_2}{(n_1-3)(n_2-2)}.$$

Thus, under synaptic fading, if there are many nodes and the assumptions of the Corollary hold then the probability of free-lunch learning is very low — in marked contrast to the behaviour under synaptic drift.

Appendix: Proofs of theorems

The main tool is the result that if $\mathbf{u}$ is uniformly distributed on S^{n-1} and $\mathbf{A}$ is an $n \times n$ matrix then

$$\mathrm{E}\left[\mathbf{u}^T\mathbf{A}\mathbf{u}\right] = \frac{\mathrm{tr}\left(\mathbf{A}\right)}{n} \qquad (A.1)$$

$$\mathrm{var}\left(\mathbf{u}^T\mathbf{A}\mathbf{u}\right) = \frac{n\mathrm{tr}\left(\mathbf{A}^2\right) + n\mathrm{tr}\left(\mathbf{A}\mathbf{A}^T\right) - 2\mathrm{tr}\left(\mathbf{A}\right)^2}{n^2(n+2)}. \qquad (A.2)$$

See (9.6.1)–(9.6.2) of Mardia and Jupp (2000).

Proof of Theorem 2. (i). Taking the conditional expectation of (2.1) over $\mathbf{v}$ and using (A.1) gives the result.

(ii). The proof is a simple extension of that of (A.26) of Stone and Jupp (2007). □

Proof of Theorem 3. (i). Taking the conditional expectation and variance of (2.1) over $\|\mathbf{v}\|^{-1}\mathbf{v}$ and using (A.1) and (A.2) gives

$$\mathrm{E}\left[\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \,|\mathbf{X}_1, \mathbf{Z}_2, \|\mathbf{v}\|\right] = \frac{\|\mathbf{v}\|^2 \mathrm{tr}\left(\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_2^T\right)}{n} \tag{A.3}$$

and

$$\begin{aligned}\mathrm{var}\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \,|\mathbf{X}_1, \mathbf{Z}_2, \|\mathbf{v}\|\right) = \frac{\|\mathbf{v}\|^4}{n(n+2)} \Big\{ & 4\mathrm{tr}\left(\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_2^T\right) \\ & - 2\mathrm{tr}\left(\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_2^T\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_2^T\right) \\ & - \frac{2}{n}\left[\mathrm{tr}\left(\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_2^T\right)\right]^2 \Big\}.\end{aligned}$$

Then Chebyshev's inequality gives

$$\begin{aligned}& P\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) < 0 \,|\mathbf{X}_1, \mathbf{Z}_2, \|\mathbf{v}\|\right) \\ & \quad \leq \frac{4\mathrm{tr}\left(\mathbf{Z}_2^T\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_2^T\mathbf{Z}_2\right)}{\left[\mathrm{tr}\left(\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_2^T\right)\right]^2}.\end{aligned} \tag{A.4}$$

Since (a) $\mathbf{Z}_2^T\mathbf{Z}_2$ is a projection operator of rank n_2 and (b) for any positive-definite $p \times p$ matrix $\mathbf{A}$ and any $p \times p$ projection matrix $\mathbf{\Pi}$ of rank r,

$$\mathrm{tr}\left(\mathbf{\Pi}\mathbf{A}^2\mathbf{\Pi}\right) \leq \frac{\left[\mathrm{tr}\left(\mathbf{\Pi}\mathbf{A}\mathbf{\Pi}\right)\right]^2}{r},$$

(A.4) gives

$$P\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) < 0 \,|\mathbf{X}_1, \mathbf{Z}_2, \|\mathbf{v}\|\right) \leq \frac{4}{n_2},$$

from which the result follows.

(ii). The proof is a simple extension of that in §A.5.2 of Stone and Jupp (2007). □

Proof of Theorem 4. Taking the conditional expectation of (4.2) over $\mathbf{d}_1$ gives (4.3). □

Proof of Theorem 5. Taking the conditional variance of (4.2) over $\mathbf{d}_1$ and using (A.2) gives

$$\begin{aligned}& \mathrm{var}\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \,|\mathbf{X}_1, \mathbf{X}_2, \mathbf{d}_2, r\right) \\ & \quad = 4(1-r)^4 \frac{\mathrm{E}\left[\|\mathbf{d}_1\|^2\right]}{n_1} \|\mathbf{X}_1\mathbf{Z}_2^T\mathbf{T}_2^{-1}\mathbf{d}_2\|^2.\end{aligned} \tag{A.5}$$

Chebyshev's inequality, (4.3) and (A.5) yield

$$P\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0 | \mathbf{X}_1, \mathbf{X}_2, \mathbf{d}_2, r\right) \leq \frac{4\mathrm{E}\left[\|\mathbf{d}_1\|^2\right]}{n_1\|\mathbf{X}_1\mathbf{a}_2\|^2}, \tag{A.6}$$

where

$$\mathbf{a}_2 = \mathbf{Z}_2^T\mathbf{T}_2^{-1}\mathbf{d}_2.$$

Since $\|\mathbf{X}_1\mathbf{a}_2\|^2 = \sum_{i=1}^{n_1} v_i^2$, where $v_i = \mathbf{x}_i{}^T\mathbf{a}_2$ for $i = 1, \ldots, n_1$ and $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ are the columns of $\mathbf{X}_1^T$, application of Jensen's inequality to the convex function $x \mapsto x^{-1}$ yields

$$\|\mathbf{X}_1\mathbf{a}_2\|^{-2} \leq \frac{1}{m^2} \sum_{j=1}^{m} \frac{1}{v_{3j-2}^2 + v_{3j-1}^2 + v_{3j}^2},$$

where $m = [n_1/3]$, and so

$$\mathrm{E}\left[\, \|\mathbf{X}_1\mathbf{a}_2\|^{-2} \big|\, \mathbf{a}_2 \right] \leq \frac{K}{\|\mathbf{a}_2\|^2 m \mathrm{E}\left[\|\mathbf{x}_i\|^2\right]}. \tag{A.7}$$

Since the density of $\mathbf{x}_i$ is bounded above, comparison of K with $\mathrm{E}\left[U^{-1}\right]$, where $U \sim \chi_3^2$, shows that K is finite.

Using the facts that $\mathbf{Z}_2\mathbf{Z}_2^T = \mathbf{I}_{n_2}$, $\mathbf{T}_2\mathbf{T}_2^T = \mathbf{X}_2\mathbf{X}_2^T = \sum_{i=1}^{n_2} \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T$, where $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{n_2}$ are the columns of $\mathbf{X}_2{}^T$, and $\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right)\left(\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}\right) \geq \|\mathbf{x}\|^4$ for any vector $\mathbf{x}$ and any positive-definite symmetric matrix $\mathbf{A}$ yields

$$\mathrm{E}\left[\|\mathbf{a}_2\|^{-2}\right] \leq \mathrm{E}\left[\|\tilde{\mathbf{x}}_i\|^2\right] \mathrm{E}\left[\|\mathbf{d}_2\|^{-2}\right]. \tag{A.8}$$

Combining (A.6)–(A.8) gives (4.4). □

References

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.

Mardia, K. V. and Jupp, P. E. (2000), *Directional Statistics*, Wiley, Chichester.

Müller, B. and Reinhardt, M. (1990), *Neural Networks*, Springer-Verlag, Berlin.

Stone, J. V. (2007), *Distributed representations accelerate evolution of adaptive behaviours*, PLoS Computational Biology **3**(8): e417.

Stone, J. V., Hunkin, N. and Hornby, A. (2001), *Predicting spontaneous recovery of memory*, Nature **414**, 167–168.

Stone, J. V. and Jupp, P. E. (2007), *Free-Lunch Learning: Modeling spontaneous recovery of memory*, Neural Computation **19**, 194–217.

Stone, J. V. and Jupp, P. E. (2008), *Falling towards forgetfulness: Synaptic decay prevents spontaneous recovery of memory.* PLoS Computational Biology **4**(8): e1000143.

Titterington, D. M. (1999), *Neural Networks*; In: Encyclopedia of Statistical Sciences. Update Volume **3** (Eds. S. Kotz, C. B. Read, and D. L. Banks), John Wiley & Sons, New York, pp. 528–535.

University of St Andrews, United Kingdom
E-mail address: pej@st-andrews.ac.uk

University of Sheffield, United Kingdom
E-mail address: j.v.stone@sheffield.ac.uk