Supplementary Material

7 Proof of Theorem 1

In this section, we prove Theorem 1 on the effect of shift invariance on the margin of a linear classifier. We call a linear classifier shift invariant when it places all possible shifts of a signal in the same class. We prove that for a shift invariant linear classifier the margin will depend only on differences in the DC components of the training signals. It follows that for the two classes shown in Figure 1, the margin of a linear, shift invariant classifier will shrink in proportion to $\frac{1}{\sqrt{d}}$, where d is the number of image pixels.

Theorem 1. *Let S_1 and S_2 denote the sets of all shifts of X_1 and X_2, as described above. They are linearly separable if and only if $\max_{\mathbf{x_1}\in S_1} f_{dc}(\mathbf{x_1}) < \min_{\mathbf{x_2}\in S_2} f_{dc}(\mathbf{x_2})$ or $\max_{\mathbf{x_2}\in S_2} f_{dc}(\mathbf{x_2}) < \min_{\mathbf{x_1}\in S_1} f_{dc}(\mathbf{x_1})$. Furthermore, if the two classes are linearly separable then, if the first inequality holds, the margin is $\min_{\mathbf{x_2}\in S_2} f_{dc}(\mathbf{x}_2) - \max_{\mathbf{x_1}\in S_1} f_{dc}(\mathbf{x}_1)$, and similarly if the second inequality holds. Furthermore, the max margin separating hyperplane has a normal of $\bar{\mathbf{w}}$.*

Proof. We only consider the case in which $\max_{\mathbf{x}_1\in S_1} f_{dc}(\mathbf{x}_1) < \min_{\mathbf{x}_2\in S_2} f_{dc}(\mathbf{x}_2)$, without loss of generality. Two classes are linearly separable iff there exists a d-dimensional unit vector $\mathbf{w}$ and a threshold T such that: $\forall \mathbf{x}_1 \in S_1, \mathbf{w}^T\mathbf{x}_1 \leq T$ and $\forall \mathbf{x}_2 \in S_2, \mathbf{w}^T\mathbf{x}_2 \geq T$. We say the margin is: $\min_{\mathbf{x}_1\in S_1, \mathbf{x}_2\in S_2} \mathbf{w}^T\mathbf{x}_2 - \mathbf{w}^T\mathbf{x}_1$.

First, it is obvious that if $\max_{\mathbf{x_1}\in S_1} f_{dc}(\mathbf{x}_1) < \min_{\mathbf{x_2}\in S_2} f_{dc}(\mathbf{x}_2)$ then S_1 and S_2 are linearly separable, by letting $\mathbf{w} = \bar{\mathbf{w}}$, and

$$T = \max_{\mathbf{x_1}\in S_1} f_{dc}(\mathbf{x}_1) + \frac{\min_{\mathbf{x_2}\in S_2} f_{dc}(\mathbf{x}_2) - \max_{\mathbf{x_1}\in S_1} f_{dc}(\mathbf{x}_1)}{2}$$

The margin for $\bar{\mathbf{w}}$ is $\min_{\mathbf{x_2}\in S_2} f_{dc}(\mathbf{x}_2) - \max_{\mathbf{x_1}\in S_1} f_{dc}(\mathbf{x}_1)$.

Second, we consider the opposite direction, supposing that the two classes are linearly separable. In this case there exist some $\mathbf{w}$ and T that will separate the classes. We can assume WLOG that $\|\mathbf{w}\| = 1$ and $f_{dc}(\mathbf{w}) > 0$. Let $\mathbf{x}_1 \in X_1$. Then we have WLOG: $\forall s \quad \mathbf{w}^T\mathbf{x}_1^s < T$. This gives:

$$\frac{\sum_{s=0}^{d-1} \mathbf{w}^T\mathbf{x}_1^s}{d} < T \implies \frac{\mathbf{w}^T \sum_{s=0}^{d-1} \mathbf{x}_1^s}{d} < T$$

Note that $\forall \mathbf{x} \in \mathbb{R}^d$, $\sum_{s=0}^{d-1} \mathbf{x}^s$ is just a constant vector of length d with each term equal to $\sqrt{d} f_{dc}(\mathbf{x})$. Therefore:

$$\frac{\sum_{s=0}^{d-1} \mathbf{w}^T\mathbf{x}^s}{d} = f_{dc}(\mathbf{w}) f_{dc}(\mathbf{x}) \tag{1}$$

so

$$f_{dc}(\mathbf{w}) f_{dc}(\mathbf{x}_1) < T \quad \forall \mathbf{x}_1 \in X_1$$

By similar reasoning, we have:

$$f_{dc}(\mathbf{w}) f_{dc}(\mathbf{x}_2) > T \quad \forall \mathbf{x}_2 \in X_2$$

Because $f_{dc}(\mathbf{w}) > 0$, $f_{dc}(\mathbf{x}_1) < f_{dc}(\mathbf{x}_2)$. So:

$$\max_{\mathbf{x}_1\in X_1} f_{dc}(\mathbf{x}_1) < \frac{T}{f_{dc}(\mathbf{w})} \quad \text{and} \quad \min_{\mathbf{x}_2\in X_2} f_{dc}(\mathbf{x}_2) > \frac{T}{f_{dc}(\mathbf{w})}$$

so

$$\max_{x_1\in X_1} f_{dc}(\mathbf{x}_1) < \min_{x_2\in X_2} f_{dc}(\mathbf{x}_2) \tag{2}$$

This shows that S_1 and S_2 are linearly separable if and only if their DC components are separable.

We now show that if S_1 and S_2 are linearly separable, for the max margin separator we have $\mathbf{w} = \bar{\mathbf{w}}$, with a margin of $\min_{\mathbf{x}_1\in S_2} f_{dc}(\mathbf{x}_2) - \max_{\mathbf{x}_1\in S_1} f_{dc}(\mathbf{x}_1)$. Because $\bar{\mathbf{w}}^T\mathbf{x} = f_{dc}(\mathbf{x})$, and the DC components are separable, $\bar{\mathbf{w}}$ separates the data, and we can see that the margin will be $\min_{\mathbf{x}\in S_2} f_{dc}(\mathbf{x}) - \max_{\mathbf{x}\in S_1} f_{dc}(\mathbf{x})$.

So, it remains to show that no other choice of $\mathbf{w}$ separates the data with a larger margin. Because we assume, WLOG that $\|\mathbf{w}\| = 1$ the margin is:

$$\min_{\mathbf{x}_2 \in S_2} \mathbf{w}^T \mathbf{x}_2 - \max_{\mathbf{x}_1 \in S_1} \mathbf{w}^T \mathbf{x}_1$$

We will show that $\forall \mathbf{x}_1 \in X_1, \mathbf{x}_2 \in X_2$ and $\forall \mathbf{w}$ such that $\|\mathbf{w}\| = 1$:

$$\min_s \mathbf{w}^T \mathbf{x}_2^s - \max_s \mathbf{w}^T \mathbf{x}_1^s \leq \min_s \bar{\mathbf{w}}^T \mathbf{x}_2^s - \max_s \bar{\mathbf{w}}^T \mathbf{x}_1^s \tag{3}$$

which implies that the margin from $\bar{\mathbf{w}}$ is greater than or equal to the margin from $\mathbf{w}$. We note that $\min_s \bar{\mathbf{w}}^T \mathbf{x}_2^s - \max_s \bar{\mathbf{w}}^T \mathbf{x}_1^s = f_{dc}(\mathbf{x}_2) - f_{dc}(\mathbf{x}_1)$.

From Equation 1 we know that

$$\max_s \mathbf{w}^T \mathbf{x}^s \geq f_{dc}(\mathbf{w}) f_{dc}(\mathbf{x}), \qquad \min_s \mathbf{w}^T \mathbf{x}^s \leq f_{dc}(\mathbf{w}) f_{dc}(\mathbf{x})$$

This implies that:

$$\min_s \mathbf{w}^T \mathbf{x}_2 - \max_s \mathbf{w}^T \mathbf{x}_1 \leq f_{dc}(\mathbf{w})(f_{dc}(\mathbf{x}_2) - f_{dc}(\mathbf{x}_1))$$

Given the constraint that $\|\mathbf{w}\| = 1$, $f_{dc}(\mathbf{w})$ is maximized by $\bar{\mathbf{w}}$, and so $f_{dc}(\mathbf{w}) \leq 1$, and

$$f_{dc}(\mathbf{w})(f_{dc}(\mathbf{x}_2) - f_{dc}(\mathbf{x}_1)) \leq f_{dc}(\mathbf{x}_2) - f_{dc}(\mathbf{x}_1)$$

and Eq. 3 is shown to hold.

□

8 Proof of Theorem 2 and Lemma 1

In this section, we first define FC networks with the neural tangent kernel (NTK) (Jacot et al., 2018) and CNNs with a convolutional neural tangent kernel (CNTK) (Arora et al., 2019; Li et al., 2019) which we will use in the proof. Then we give the proofs of Theorem 2 and Lemma 1.

Let $\mathbf{x} \in \mathbb{R}^d$ denote the input to the network. A two-layer fully connected network is defined by

$$f_{\mathrm{FC}}(\mathbf{x}; W, \mathbf{v}) = \mathbf{v}^T \sigma(W\mathbf{x}), \tag{4}$$

where $W \in \mathbb{R}^{m \times d}$ and $\mathbf{v} \in \mathbb{R}^m$ are learnable parameters and $\sigma(.)$ is the ReLU function applied elementwise. Assuming W and $\mathbf{v}$ are initialized with normal distribution, the corresponding FC-NTK for inputs $\mathbf{z}, \mathbf{x} \in \mathbb{R}^d$ is given by (Bietti & Mairal, 2019)

$$k(\mathbf{z}, \mathbf{x}) = \frac{1}{\pi}\left(2\mathbf{z}^T\mathbf{x}(\pi - \phi) + \|\mathbf{z}\|\|\mathbf{x}\| \sin\phi\right), \tag{5}$$

where ϕ denotes the angle between $\mathbf{z}$ and $\mathbf{x}$, i.e., $\phi = \arccos\left(\frac{\mathbf{z}^T\mathbf{x}}{\|\mathbf{z}\|\|\mathbf{x}\|}\right)$.

Next we define the shift invariant convolutional model. Given an input $\mathbf{x} \in \mathbb{R}^d$ and filters $\{\mathbf{w}_i\}_{i=1}^m \subset \mathbb{R}^q$ we denote by $\mathbf{w}_i * \mathbf{x} \in \mathbb{R}^d$ the circular convolution of $\mathbf{x}$ with the filter $\mathbf{w}_i$ (with no bias). $W * \mathbf{x}$ denotes the results of these convolutions, represented as an $m \times d$ matrix, with the $m \times q$ matrix W denoting the collection of all filters $\{\mathbf{w}_i\}_{i=1}^m$. Finally, let $\mathbf{v} \in \mathbb{R}^m$. Then a two layer convolutional network with global average pooling is defined by

$$f_{\mathrm{Conv}}(\mathbf{x}; W, \mathbf{v}) = \frac{1}{d}\mathbf{v}^T \sigma(W * \mathbf{x}))\mathbf{1}_d, \tag{6}$$

W and $\mathbf{v}$ include the learnable parameters initialized with the standard normal distribution and $\mathbf{1}_d \in \mathbb{R}^d$ is the vector of all ones. In this model the input $\mathbf{x}$ is convolved with the rows of W. After ReLU the result undergoes a 1×1 convolution with parameters $\mathbf{v}$ followed by global average pooling, captured by the multiplication with $\frac{1}{d}\mathbf{1}_d$.

Given inputs $\mathbf{z}, \mathbf{x} \in \mathbb{R}^d$, denote by $\bar{\mathbf{z}}_i, \bar{\mathbf{x}}_j \in \mathbb{R}^q$ their (cyclic) patches, $1 \leq i, j \leq d$, so for example $\bar{\mathbf{z}}_i = (z_i, z_{(i+1) \bmod d}, \ldots z_{(i+q-1) \bmod d})^T$. Then the corresponding CNTK-GAP $K(\mathbf{z}, \mathbf{x})$ is constructed as follows.

$$K(\mathbf{z}, \mathbf{x}) = \frac{1}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d} k(\bar{\mathbf{z}}_i, \bar{\mathbf{x}}_j), \tag{7}$$

where $k(\bar{\mathbf{z}}_i, \bar{\mathbf{x}}_j)$ is the FC-NTK given by (5) (see a related construction in (Tachella et al., 2020)).

We use FC-NTK and CNTK-GAP in kernel regression. Given training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathbb{R}$, kernel ridge regression is the solution to (Saitoh & Sawano, 2016)

$$g_k = \arg\min_{g \in \mathcal{H}_k} \sum_{i=1}^{n} (g(\mathbf{x}_i) - y_i)^2 + \lambda \|g\|^2_{\mathcal{H}_k}, \tag{8}$$

where $\mathcal{H}_k$ denotes the reproducing kernel Hilbert space associated with k. The solution of (8) is given by

$$g_k(\mathbf{z}) = (k(\mathbf{z}, \mathbf{x}_1), ..., k(\mathbf{z}, \mathbf{x}_n))(H_k + \lambda I)^{-1}\mathbf{y}, \tag{9}$$

where H_k is the $n \times n$ matrix with its i, j'th entry $k(\mathbf{x}_i, \mathbf{x}_j)$, I denotes the identity matrix, and $\mathbf{y} = (y_1, ..., y_n)^T$. Below we consider the minimum norm interpolant, i.e.,

$$g_k = \arg\min_{g \in \mathcal{H}_k} \|g\|_{\mathcal{H}_k} \quad \text{s.t.} \quad \forall i,\ g(\mathbf{x}_i) = y_i. \tag{10}$$

which is obtained when we let $\lambda \to 0$.

Theorem 2. *Let $\mathbf{x}, -\mathbf{x} \in \mathbb{R}^d$ be two training vectors with class labels $1, -1$ respectively.*

1. *Let $k(\mathbf{z}, \mathbf{x})$ denote NTK for the bias-free, two-layer fully connected network. Then $\forall \mathbf{z} \in \mathbb{R}^d$, the minimum norm interpolant $g_k(\mathbf{z}) \geq 0$ iff $\mathbf{z}^T\mathbf{x} \geq 0$.*
2. *Let $K(\mathbf{z}, \mathbf{x})$ denote CNTK-GAP for the bias-free, two-layer convolutional network, and assume H_K is invertible. Then $\forall \mathbf{z} \in \mathbb{R}^d$, either $g_K(\mathbf{z}) \geq 0$ iff $\mathbf{z}^T\mathbf{1}_d \geq 0$ or $g_K(\mathbf{z}) \geq 0$ iff $\mathbf{z}^T\mathbf{1}_d \leq 0$. (I.e., $\mathbf{z}^T\mathbf{1}_d = 0$ forms a separating hyperplane.)*

The theorem tells us that NTK and CNTK produce linear classifiers. (1) tells us that NTK produces a separating hyperplane with a normal vector $\mathbf{x}$, while (2) says that for CNTK the normal direction is $\mathbf{1}_d$.

Proof. 1. Solving the regression problem (9) with $\lambda \to 0$ we have

$$H_k = \begin{pmatrix} k(\mathbf{x}, \mathbf{x}) & k(\mathbf{x}, -\mathbf{x}) \\ k(-\mathbf{x}, \mathbf{x}) & k(-\mathbf{x}, -\mathbf{x}) \end{pmatrix} = 2\mathbf{x}^T\mathbf{x}I,$$

where the latter equality is obtained from (5) by noting that $\phi = 0$ along the diagonal and $\phi = \pi$ for the off-diagonal entries. Therefore, using (9) and noting that $\mathbf{y} = (1, -1)^T$,

$$g_k(\mathbf{z}) = \frac{1}{2x^Tx}(k(\mathbf{z}, \mathbf{x}) - k(\mathbf{z}, -\mathbf{x}))$$

Given a test point $\mathbf{z} \in \mathbb{R}^d$, let ϕ now denote the angle between $\mathbf{z}$ and $\mathbf{x}$ and note that the angle between $\mathbf{z}$ and $-\mathbf{x}$ is $\pi - \phi$. Therefore,

$$\begin{aligned} k(\mathbf{z}, \mathbf{x}) &= \frac{1}{\pi}\left(2\mathbf{z}^T\mathbf{x}(\pi - \phi) + \|\mathbf{z}\|\|\mathbf{x}\|\sin\phi\right) \\ k(\mathbf{z}, -\mathbf{x}) &= \frac{1}{\pi}\left(-2\mathbf{z}^T\mathbf{x}\phi + \|\mathbf{z}\|\|\mathbf{x}\|\sin\phi\right), \end{aligned}$$

implying that

$$k(\mathbf{z}, \mathbf{x}) - k(\mathbf{z}, -\mathbf{x}) = 2\mathbf{z}^T\mathbf{x}. \tag{11}$$

from which we obtain $g_k(\mathbf{z}) = \frac{\mathbf{z}^T\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$. Consequently, $g_k(\mathbf{z}) > 0$ if and only if $\mathbf{z}^T\mathbf{x} > 0$.

2. Using the definition of K (Eq. 7) it is clear that $K(\mathbf{x}, \mathbf{x}) = K(-\mathbf{x}, -\mathbf{x})$. Therefore, using Lemma 1 we need to show that $K(\mathbf{z}, \mathbf{x}) > K(\mathbf{z}, -\mathbf{x})$ on one side of the plane $\mathbf{z}^T\mathbf{1}_d = 0$. Consider the patches $\bar{\mathbf{z}}_i$ in $\mathbf{z}$ and $\bar{\mathbf{x}}_j$ in $\mathbf{x}$. From (11) we have

$$k(\bar{\mathbf{z}}_i, \bar{\mathbf{x}}_j) - k(\bar{\mathbf{z}}_i, -\bar{\mathbf{x}}_j) = 2\bar{\mathbf{z}}_i^T\bar{\mathbf{x}}_j,$$

implying that

$$K(\mathbf{z}, \mathbf{x}) - K(\mathbf{z}, -\mathbf{x}) = \frac{2}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d}\bar{\mathbf{z}}_i^T\bar{\mathbf{x}}_j.$$

Rewriting this in matrix notation we have

$$K(\mathbf{z},\mathbf{x}) - K(\mathbf{z},-\mathbf{x}) = \frac{2}{d^2}\mathbf{1}_d^T Z^T X \mathbf{1}_d,$$

where Z and X are $q \times d$ matrices whose columns respectively contain all the patches of $\mathbf{z}$ and $\mathbf{x}$. Since all rows of Z and X are identical up to a cyclic permutation $\hat{z} = Z\mathbf{1}_d$ and $\hat{x} = X\mathbf{1}_d$ are vectors of constants in $\mathbb{R}^q$ with the constants $\mathbf{z}^T\mathbf{1}_d$ and $\mathbf{x}^T\mathbf{1}_d$ respectively. Consequently, using Lemma 1

$$g_K(\mathbf{z}) = c(K(\mathbf{z},\mathbf{x}) - K(\mathbf{z},-\mathbf{x})) = \frac{2cq}{d^2}(\mathbf{z}^T\mathbf{1}_d)(\mathbf{x}^T\mathbf{1}_d).$$

where, because K is positive definite and H_K is invertible, $c = 1/(K(\mathbf{x},\mathbf{x}) - K(\mathbf{x},-\mathbf{x})) > 0$. Denoting $\beta = \frac{2cq}{d^2}(\mathbf{x}^T\mathbf{1}_d)$, we obtain that $g_K(\mathbf{z}) > 0$ if and only if $\operatorname{sign}(\beta)\,\mathbf{z}^T\mathbf{1}_d > 0$, proving the theorem.

□

The following lemma was used to prove Thm. 2.

Lemma 1. *Let $k(.,.)$ be a positive definite kernel with a training set $\{(\mathbf{x}_1,+1),(\mathbf{x}_2,-1)\} \subset \mathbb{R}^d \times \mathbb{R}$. If $k(\mathbf{x}_1,\mathbf{x}_1) = k(\mathbf{x}_2,\mathbf{x}_2)$ and H_k is invertible with $\lambda \to 0$ then a test point $\mathbf{z} \in \mathbb{R}^d$ is classified as +1 if and only $k(\mathbf{z},\mathbf{x}_1) > k(\mathbf{z},\mathbf{x}_2)$.*

Proof. Denote by $a = k(\mathbf{x}_1,\mathbf{x}_1) = k(\mathbf{x}_2,\mathbf{x}_2)$ and $b = k(\mathbf{x}_1,\mathbf{x}_2)$, then $H_k = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$. Clearly, $\mathbf{y} = (1,-1)^T$ is an eigenvector of H_k with the eigenvalue $a - b > 0$, which is positive due to the positive definiteness of k. Consequently, $\mathbf{y}$ is also an eigenvector of H_k^{-1} with eigenvalue $1/(a-b) > 0$. Applying (9) we have

$$\begin{aligned} g_k(\mathbf{z}) &= (k(\mathbf{z},\mathbf{x}_1), k(\mathbf{z},\mathbf{x}_2))H^{-1}\mathbf{y} \\ &= \frac{1}{a-b}(k(\mathbf{z},\mathbf{x}_1) - k(\mathbf{z},\mathbf{x}_2)) \end{aligned}$$

Therefore, $g_k(\mathbf{z}) > 0$ if and only if $k(\mathbf{z},\mathbf{x}_1) > k(\mathbf{z},\mathbf{x}_2)$. □