Supplementary Materials

A Inverse polynomial kernel and Gaussian kernel

In this appendix, we describe the properties of the two types of kernels — the inverse polynomial kernel (14) and the Gaussian RBF kernel (15). We prove that the associated reproducing kernel Hilbert Spaces (RKHS) of these kernels contain filters taking the form $h : z \mapsto \sigma(\langle w, z\rangle)$ for particular activation functions σ.

A.1 Inverse polynomial kernel

We first verify that the function (14) is a kernel function. This holds since that we can find a mapping $\varphi : \mathbb{R}^{d_1} \to \ell^2(\mathbb{N})$ such that $\mathcal{K}(z, z') = \langle \varphi(z), \varphi(z')\rangle$. We use z_i to represent the i-th coordinate of an infinite-dimensional vector z. The $(k_1, \ldots, k_j)$-th coordinate of $\varphi(z)$, where $j \in \mathbb{N}$ and $k_1, \ldots, k_j \in [d_1]$, is defined as $2^{-\frac{j+1}{2}} x_{k_1} \ldots x_{k_j}$. By this definition, we have

$$\langle \varphi(x), \varphi(y)\rangle = \sum_{j=0}^{\infty} 2^{-(j+1)} \sum_{(k_1,\ldots,k_j)\in[d_1]^j} z_{k_1} \ldots z_{k_j} z'_{k_1} \ldots z'_{k_j}. \tag{21}$$

Since $\|z\|_2 \le 1$ and $\|z'\|_2 \le 1$, the series on the right-hand side is absolutely convergent. The inner term on the right-hand side of equation (21) can be simplified to

$$\sum_{(k_1,\ldots,k_j)\in[d_1]^j} z_{k_1} \ldots z_{k_j} z'_{k_1} \ldots z'_{k_j} = (\langle z, z'\rangle)^j. \tag{22}$$

Combining equations (21) and (22) and using the fact that $|\langle z, z'\rangle| \le 1$, we have

$$\langle \varphi(z), \varphi(z')\rangle = \sum_{j=0}^{\infty} 2^{-(j+1)} (\langle z, z'\rangle)^j \overset{\text{(i)}}{=} \frac{1}{2 - \langle z, z'\rangle} = \mathcal{K}(z, z'),$$

which verifies that $\mathcal{K}$ is a kernel function and φ is the associated feature map. Next, we prove that the associated RKHS contains the class of nonlinear filters. The lemma was proved by Zhang et al. [2]. We include the proof to make the paper self-contained.

Lemma 1. *Assume that the function $\sigma(x)$ has a polynomial expansion $\sigma(t) = \sum_{j=0}^{\infty} a_j t^j$. Let $C_\sigma(\lambda) := \sqrt{\sum_{j=0}^{\infty} 2^{j+1} a_j^2 \lambda^{2j}}$. If $C_\sigma(\|w\|_2) < \infty$, then the RKHS induced by the inverse polynomial kernel contains function $h : z \mapsto \sigma(\langle w, z\rangle)$ with Hilbert norm $\|h\|_{\mathcal{H}} = C_\sigma(\|w\|_2)$.*

Proof. Let φ be the feature map that we have defined for the polynomial inverse kernel. We define vector $\overline{w} \in \ell^2(\mathbb{N})$ as follow: the $(k_1, \ldots, k_j)$-th coordinate of $\overline{w}$, where $j \in \mathbb{N}$ and $k_1, \ldots, k_j \in [d_1]$, is equal to $2^{\frac{j+1}{2}} a_j w_{k_1} \ldots w_{k_j}$. By this definition, we have

$$\sigma(\langle w, z \rangle) = \sum_{t=0}^{\infty} a_j (\langle w, z \rangle)^j = \sum_{j=0}^{\infty} a_j \sum_{(k_1,\ldots,k_j) \in [d_1]^j} w_{k_1} \ldots w_{k_j} z_{k_1} \ldots z_{k_j} = \langle \overline{w}, \varphi(z) \rangle, \tag{23}$$

where the first equation holds since $\sigma(x)$ has a polynomial expansion $\sigma(x) = \sum_{j=0}^{\infty} a_j x^j$, the second by expanding the inner product, and the third by definition of $\overline{w}$ and $\varphi(z)$. The ℓ_2-norm of $\overline{w}$ is equal to:

$$\|\overline{w}\|_2^2 = \sum_{j=0}^{\infty} 2^{j+1} a_j^2 \sum_{(k_1,\ldots,k_j) \in [d_1]^j} \overline{w}_{k_1}^2 \overline{w}_{k_2}^2 \cdots \overline{w}_{k_j}^2 = \sum_{j=0}^{\infty} 2^{j+1} a_j^2 \|\overline{w}\|_2^{2j} = C_\sigma^2(\|\overline{w}\|_2) < \infty. \tag{24}$$

By the basic property of the RKHS, the Hilbert norm of h is equal to the ℓ_2-norm of $\overline{w}$. Combining equations (23) and (24), we conclude that $h \in \mathcal{H}$ and $\|h\|_{\mathcal{H}} = \|\overline{w}\|_2 = C_\sigma(\|\overline{w}\|_2)$. □

According to Lemma 1, it suffices to upper bound $C_\sigma(\lambda)$ for a particular activation function σ. To make $C_\sigma(\lambda) < \infty$, the coefficients $\{a_j\}_{j=0}^{\infty}$ must quickly converge to zero, meaning that the activation function must be sufficiently smooth. For polynomial functions of degree ℓ, the definition of C_σ implies that $C_\sigma(\lambda) = \mathcal{O}(\lambda^\ell)$. For the sinusoid activation $\sigma(t) := \sin(t)$, we have

$$C_\sigma(\lambda) = \sqrt{\sum_{j=0}^{\infty} \frac{2^{2j+2}}{((2j+1)!)^2} \cdot (\lambda^2)^{2j+1}} \leq 2e^{\lambda^2}.$$

For the erf function and the smoothed hinge loss function defined in Section 3.4, Zhang et al. [2, Proposition 1] proved that $C_\sigma(\lambda) = \mathcal{O}(e^{c\lambda^2})$ for universal numerical constant $c > 0$.

A.2 Gaussian kernel

The Gaussian kernel also induces an RKHS that contains a particular class of nonlinear filters. The proof is similar to that of Lemma 1.

Lemma 2. *Assume that the function $\sigma(x)$ has a polynomial expansion $\sigma(t) = \sum_{j=0}^{\infty} a_j t^j$. Let $C_\sigma(\lambda) := \sqrt{\sum_{j=0}^{\infty} \frac{j! e^{2\gamma}}{(2\gamma)^j} a_j^2 \lambda^{2j}}$. If $C_\sigma(\|w\|_2) < \infty$, then the RKHS induced by the Gaussian kernel contains the function $h : z \mapsto \sigma(\langle w, z \rangle)$ with Hilbert norm $\|h\|_{\mathcal{H}} = C_\sigma(\|w\|_2)$.*

Proof. When $\|z\|_2 = \|z'\|_2 = 1$, It is well-known [see, e.g. 1] the following mapping $\varphi : \mathbb{R}^{d_1} \to \ell^2(\mathbb{N})$ is a feature map for the Gaussian RBF kernel: the $(k_1, \ldots, k_j)$-th coordinate of $\varphi(z)$, where $j \in \mathbb{N}$ and $k_1, \ldots, k_j \in [d_1]$, is defined as $e^{-\gamma}((2\gamma)^j / j!)^{1/2} x_{k_1} \ldots x_{k_j}$. Similar to equation (23), we define a vector $\overline{w} \in \ell^2(\mathbb{N})$ as follow: the $(k_1, \ldots, k_j)$-th coordinate of $\overline{w}$, where $j \in \mathbb{N}$ and $k_1, \ldots, k_j \in [d_1]$, is equal to $e^{\gamma}((2\gamma)^j / j!)^{-1/2} a_j w_{k_1} \ldots w_{k_j}$. By this definition, we have

$$\sigma(\langle w, z \rangle) = \sum_{t=0}^{\infty} a_j (\langle w, z \rangle)^j = \sum_{j=0}^{\infty} a_j \sum_{(k_1,\ldots,k_j) \in [d_1]^j} w_{k_1} \ldots w_{k_j} z_{k_1} \ldots z_{k_j} = \langle \overline{w}, \varphi(z) \rangle. \tag{25}$$

The ℓ_2-norm of $\overline{w}$ is equal to:

$$\|\overline{w}\|_2^2 = \sum_{j=0}^{\infty} \frac{j! e^{2\gamma}}{(2\gamma)^j} a_j^2 \sum_{(k_1,\dots,k_j)\in[d_1]^j} \overline{w}_{k_1}^2 \overline{w}_{k_2}^2 \cdots \overline{w}_{k_j}^2 = \sum_{j=0}^{\infty} \frac{j! e^{2\gamma}}{(2\gamma)^j} a_j^2 \|\overline{w}\|_2^{2j} = C_\sigma^2(\|\overline{w}\|_2) < \infty. \tag{26}$$

Combining equations (23) and (24), we conclude that $h \in \mathcal{H}$ and $\|h\|_{\mathcal{H}} = \|\overline{w}\|_2 = C_\sigma(\|\overline{w}\|_2)$. □

Comparing Lemma 1 and Lemma 2, we find that the Gaussian kernel imposes a stronger condition on the smoothness of the activation function. For polynomial functions of degree ℓ, we still have $C_\sigma(\lambda) = \mathcal{O}(\lambda^\ell)$. For the sinusoid activation $\sigma(t) := \sin(t)$, it can be verified that

$$C_\sigma(\lambda) = \sqrt{e^{2\gamma} \sum_{j=0}^{\infty} \frac{1}{(2j+1)!} \cdot \left(\frac{\lambda^2}{2\gamma}\right)^{2j+1}} \le e^{\lambda^2/(4\gamma)+\gamma}.$$

However, the value of $C_\sigma(\lambda)$ is infinite when σ is the erf function or the smoothed hinge loss, meaning that the Gaussian kernel's RKHS doesn't contain filters activated by these two functions.

B Convex relaxation for nonlinear activation

In this appendix, we provide a detailed derivation of the relaxation for nonlinear activation functions that we previously sketched in Section 3.2. Recall that the filter output is $\sigma(\langle w_j, z\rangle)$. Appendix A shows that given a sufficiently smooth activation function σ, we can find some kernel function $\mathcal{K} : \mathbb{R}^{d_1} \times \mathbb{R}^{d_1} \to \mathbb{R}$ and a feature map $\varphi : \mathbb{R}^{d_1} \to \ell^2(\mathbb{N})$ satisfying $\mathcal{K}(z, z') \equiv \langle \varphi(z), \varphi(z')\rangle$, such that

$$\sigma(\langle w_j, z\rangle) \equiv \langle \overline{w}_j, \varphi(z)\rangle. \tag{27}$$

Here $\overline{w}_j \in \ell^2(\mathbb{N})$ is a countable-dimensional vector and $\varphi := (\varphi_1, \varphi_2, \dots)$ is a countable sequence of functions. Moreover, the ℓ_2-norm of $\overline{w}_j$ is bounded as $\|\overline{w}_j\|_2 \le C_\sigma(\|w_j\|_2)$ for a monotonically increasing function C_σ that depends on the kernel (see Lemma 1 and Lemma 2). As a consequence, we may use $\varphi(z)$ as the vectorized representation of the patch z, and use $\overline{w}_j$ as the linear transformation weights, then the problem is reduced to training a CNN with the identity activation function.

The filter is parametrized by an infinite-dimensional vector $\overline{w}_j$. Our next step is to reduce the original ERM problem to a finite-dimensional one. In order to minimize the empirical risk, one only needs to concern the output on the training data, that is, the output of $\langle \overline{w}_j, \varphi(z_p(x_i))\rangle$ for all $(i, p) \in [n] \times [P]$. Let T be the orthogonal projector onto the linear subspace spanned by the vectors $\{\varphi(z_p(x_i)) : (i, p) \in [n] \times [P]\}$. Then we have

$$\forall\ (i,p) \in [n] \times [P] : \quad \langle \overline{w}_j, \varphi(z_p(x_i))\rangle = \langle \overline{w}_j, T\varphi(z_p(x_i))\rangle = \langle T\overline{w}_j, \varphi(z_p(x_i))\rangle.$$

The last equation follows since the orthogonal projector T is self-adjoint. Thus, for empirical risk minimization, we can without loss of generality assume that $\overline{w}_j$ belongs to the linear subspace spanned by $\{\varphi(z_p(x_i)) : (i, p) \in [n] \times [P]\}$ and reparametrize it by:

$$\overline{w}_j = \sum_{(i,p)\in[n]\times[P]} \beta_{j,(i,p)} \varphi(z_p(x_i)). \tag{28}$$

Let $\beta_j \in \mathbb{R}^{nP}$ be a vector whose whose (i,p)-th coordinate is $\beta_{j,(i,p)}$. In order to estimate $\overline{w}_j$, it suffices to estimate the vector β_j. By definition, the vector satisfies the relation $\beta_j^\top K \beta_j = \|\overline{w}_j\|_2^2$, where K is the $nP \times nP$ kernel matrix defined in Section 3.2. As a consequence, if we can find a matrix Q such that $QQ^\top = K$, then we have the norm constraint

$$\|Q^\top \beta_j\|_2 = \sqrt{\beta_j^\top K \beta_j} = \|\overline{w}_j\|_2 \leq C_\sigma(\|w_j\|_2) \leq C_\sigma(B). \tag{29}$$

Let $v(z) \in \mathbb{R}^{nP}$ be a vector whose (i,p)-th coordinate is equal to $\mathcal{K}(z, z_p(x_i))$. Then by equations (27) and (28), the filter output can be written as

$$\sigma\big(\langle w_j,\, z\rangle\big) \equiv \langle \overline{w}_j,\, \varphi(z)\rangle \;\equiv\; \langle \beta_j,\, v(z)\rangle. \tag{30}$$

For any patch $z_p(x_i)$ in the training data, the vector $v(z_p(x_i))$ belongs to the column space of the kernel matrix K. Therefore, letting $Q^\dagger$ represent the pseudo-inverse of matrix Q, we have

$$\forall\ (i,p) \in [n] \times [P]: \quad \langle \beta_j,\, v(z_p(x_i))\rangle = \beta_j^\top Q Q^\dagger v(z_p(x_i)) = \langle (Q^\top)^\dagger Q^\top \beta_j,\, v(z_p(x_i))\rangle.$$

It means that if we replace the vector β_j on the right-hand side of equation (30) by the vector $(Q^\top)^\dagger Q^\top \beta_j$, then it won't change the empirical risk. Thus, for ERM we can parametrize the filters by

$$h_j(z) := \langle (Q^\top)^\dagger Q^\top \beta_j,\, v(z)\rangle = \langle Q^\dagger v(z),\, Q^\top \beta_j\rangle. \tag{31}$$

Let $Z(x)$ be an $P \times nP$ matrix whose p-th row is equal to $Q^\dagger v(z_p(x))$. Similar to the steps in equation (6), we have

$$f_k(x) = \sum_{j=1}^{r} \alpha_{k,j}^\top Z(x) Q^\top \beta_j = \mathrm{tr}\Big(Z(x)\Big(\sum_{j=1}^{r} Q^\top \beta_j \alpha_{k,j}^\top\Big)\Big) = \mathrm{tr}(Z(x) A_k),$$

where $A_k := \sum_{j=1}^r Q^\top \beta_j \alpha_{k,j}^\top$. If we let $A := (A_1, \ldots, A_{d_2})$ denote the concatenation of these matrices, then this larger matrix satisfies the constraints:

Constraint (C1): $\max\limits_{j \in [r]} \|Q^\top \beta_j\|_2 \leq C_\sigma(B_1)$ and $\max\limits_{(k,j) \in [d_2] \times [r]} \|\alpha_{k,j}\|_2 \leq B_2$.

Constraint (C2): The matrix A has rank at most r.

We relax these two constraints to the nuclear norm constraint:

$$\|A\|_* \leq C_\sigma(B_1) B_2 r \sqrt{d_2}. \tag{32}$$

By comparing constraints (8) and (32), we see that the only difference is that the term B_1 in the norm bound has been replaced by $C_\sigma(B_1)$. This change is needed because we have used the kernel trick to handle nonlinear activation functions.

References

[1] I. Steinwart and A. Christmann. *Support vector machines*. Springer, New York, 2008.

[2] Y. Zhang, J. D. Lee, and M. I. Jordan. ℓ_1-regularized neural networks are improperly learnable in polynomial time. In *Proceedings on the 33rd International Conference on Machine Learning*, 2016.