Inductive Bias of Multi-Channel Linear Convolutional Networks with Bounded Weight Norm

Meena Jagadeesan*[1], Ilya Razenshteyn†[2], and Suriya Gunasekar‡[3]

[1]University of California, Berkeley
[2]CipherMode Labs
[3]Microsoft Research

Abstract

We study the function space characterization of the inductive bias resulting from controlling the ℓ_2 norm of the weights in linear convolutional networks. We view this in terms of an *induced regularizer* in the function space given by the minimum norm of weights required to realize a linear function. For two layer linear convolutional networks with C output channels and kernel size K, we show the following: (a) If the inputs to the network have a single channel, the induced regularizer for any K is a norm given by a semidefinite program (SDP) that is *independent* of the number of output channels C. (b) In contrast, for networks with multi-channel inputs, multiple output channels can be necessary to merely realize all matrix-valued linear functions and thus the inductive bias *does* depend on C. Further, for sufficiently large C, the induced regularizer for $K = 1$ and $K = D$ are the nuclear norm and the $\ell_{2,1}$ group-sparse norm, respectively, of the Fourier coefficients. (c) Complementing our theoretical results, we show through experiments on MNIST and CIFAR-10 that our key findings extend to implicit biases from gradient descent in overparameterized networks.

1 Introduction

In the study of generalization and model capacity, it has long been argued that the magnitude of parameters plays a greater role in learning than the number of parameters [Bar96; BM02; NTS15; ZBHRV17; BFT17]. In the current practice of deep learning, capacity control of weight magnitudes is typically achieved through a combination of explicit regularization techniques, as well as implicit regularization from optimization algorithms like stochastic gradient descent. In particular, regularization of ℓ_2 norm of weights (closely related to weight decay) is by far the most common way to explicitly control the magnitude of weights [KH91; WLLM19]. Further, implicit bias from (stochastic) gradient descent in many overparameterized models has also been prominently connected to complexity control in terms of ℓ_2 norm of the weights (see *e.g.,* Lyu and Li [LL20] and Nacson et al. [NGLSS19]; additional references and discussions in Sections 1.1).

Controlling the ℓ_2 norm of weights while fitting the training data is thus an important inductive bias, but this description does not directly provide insight into properties of the learned function. Even for networks that realize the same model class, minimizing or bounding the ℓ_2 norm of weights

*mjagadeesan@berkeley.edu. This work was partly done while M. Jagadeesan and I. Razenshteyn were at Microsoft Research. M. Jagadeesan was supported in part by the Paul and Daisy Soros Fellowship.

†ilya.razenshteyn@gmail.com
‡suriyag@microsoft.com

in different architectures can lead to remarkably different effects in function space. For example, consider linear networks (networks with linear activation) with fully connected and convolution layers, which are simply different parameterizations of the same model class of linear functions. In a work closely related to ours, Gunasekar et al. [GLSS18b] showed for fully connected linear networks, minimizing the ℓ_2 norm of weights corresponds to minimizing the ℓ_2 norm of the linear map realized by the network, while for linear convolutional network with *full dimensional kernels*, it corresponds to minimizing the ℓ_1 norm of Fourier coefficients of the linear map. This function space view reveals that minimizing the ℓ_2 norm of weights in these networks has fundamentally different implications for learned predictors that depends on specific parametrization of function class.

The function space view of controlling ℓ_2 norm of parameters can be understood in terms of the *representation cost* with respect to the weight norm, that is, the minimum ℓ_2 norm of weights needed to realize a function using a given network architecture (see eq. 2). This defines an *induced complexity measure* over functions, which we also refer as the *induced regularizer*.

In this work, we investigate the impact of network architecture on the induced regularizer, focusing on the class of two-layer linear networks with a multi-channel convolution layer. In particular, we study the induced regularizer as a function of the kernel size K and the number of output and input channels (denoted by C and R, respectively), for networks with D-dimensional inputs. Interestingly, even within the class of two-layer linear convolutional networks, the induced regularizer can exhibit strikingly different properties for different choices of K, C, and R. For example, for single-channel networks, for kernel size $K = D$, the induced regularizer is the ℓ_1 norm of the Fourier coefficients [GLSS18b], while for $K = 1$, the induced regularizer turns out to be the ℓ_2 norm of the linear function.

While the induced regularizer takes on a simple form in these special cases, we demonstrate that the induced regularizer is unlikely to admit closed form solutions for general kernel sizes (see Lemma 3). In order to show properties of the induced regularizer, our main insight is to construct a *semidefinite program (SDP) relaxation of the induced regularizer* that turns out to be independent of the number of output channels C. We show that this SDP is tight in a number of interesting cases, which enables us to reason about the impact of C on the induced regularizer, explicitly characterize the induced regularizer in special cases, and formulate a testable hypothesis that we evaluate in simple experimental setups. Our main technical contributions are the following:

1. *Single-channel inputs.* For any kernel size K, we show that the SDP relaxation of the induced regularizer is tight (see Theorem 4), and consequently the induced regularizer over linear functions is a norm that is *independent* of the number of output channels C. Thus, as we increase the kernel size K, the induced regularizer interpolates between the ℓ_2 norm (for $K = 1$) and the ℓ_1 norm of the Fourier transform (for $K = D$), with larger kernels promoting sparsity in the frequency domain.
2. *Multi-channel inputs.* For multi-channel inputs of dimensions $D \times R$, even realizing all linear functions over the inputs can require multiple output channels C (see Lemma 9). Hence, the induced regularizer *does* depend on C, although for large enough C, we show that the SDP relaxation is tight. Using the SDP, we characterize the induced regularizer in the special cases of $K = 1$ and $K = D$ as the nuclear norm and the $\ell_{2,1}$ group sparse norm of the Fourier coefficients, respectively (see Theorems 12-13). In both cases, the induced regularizers promote non-trivial sparse structures.
3. *Experiments for gradient descent.* Although we do not directly study gradient descent, our results (when combined with prior work discussed in Section 1.1) suggest hypotheses about the extent to which models learned with implicit regularization from gradient descent might depend on the number of output channels. We provide empirical evidence for these hypotheses on MNIST and CIFAR-10 datasets (see Section 6).

4. *Proof technique.* In the proof of Theorem 4 on SDP tightness, we show an interesting property about convolutions involving smaller kernels of size $K < D$ (Lemma 5). To our knowledge, this result and the proof technique are new and are of independent interest.

1.1 Connections to implicit regularization

The inductive bias of controlling the ℓ_2 norm of weights or parameters is in part motivated by the recent findings about the implicit bias introduced by gradient descent–like algorithms. In many (but not all) instances of overparameterized models, it has been shown that gradient descent updates tend towards solutions that implicitly control the ℓ_2 of the parameters. To provide context for this discussion, we state the main result from Lyu and Li [LL20] on the asymptotic behaviour of gradient flow (infinitesimal step-size gradient descent) on positive homogeneous model classes, which includes networks with linear and ReLU activations. Consider a function class represented as $\Phi(\boldsymbol{\theta}; \mathbf{x})$ for inputs denoted as $\mathbf{x}$ and parameters (or weights) denoted as $\boldsymbol{\theta}$.

Theorem. *[Paraphrased from LL20] Assume that Φ is locally Lipschitz and positive homogeneous with order $L > 0$, i.e., $\forall_{\boldsymbol{\theta},\alpha>0}$, $\Phi(\alpha\boldsymbol{\theta}; .) = \alpha^L \Phi(\boldsymbol{\theta}; .)$. Consider minimizing an exponential-tailed loss over a separable binary classification dataset $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$. Under assumptions of loss convergence, gradient flow on this loss converges in direction to a first order stationary point (KKT point) of the following max–ℓ_2 margin problem in parameter space:*

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_2^2 \;\; s.t., \;\; \forall_n y_n \Phi(\boldsymbol{\theta}; \mathbf{x}_n) \geq 1. \tag{1}$$

It is easy to see that in the function space, we can define an induced regularizer $\mathcal{R}_\Phi$ such that the minimizers of the *max–ℓ_2 margin* problem over parameters in eq. (1) are equivalent to the minimizers of following *max–$\mathcal{R}_\Phi$ margin* problem over functions: $\min_f \mathcal{R}_\Phi(f)$ s.t., $\forall_n y_n f(\mathbf{x}_n) \geq 1$. The induced regularizer $\mathcal{R}_\Phi$ corresponding to controlling ℓ_2 norm of parameters is given as follows:

$$\mathcal{R}_\Phi(f) := \inf_{\boldsymbol{\theta}} \;\; \|\boldsymbol{\theta}\|_2^2 \;\; \text{s.t.}, \;\; \forall \mathbf{x}, f(\mathbf{x}) = \Phi(\boldsymbol{\theta}, \mathbf{x}). \tag{2}$$

We thus expect the implicit bias from gradient descent to be related to $\mathcal{R}_\Phi(f)$. However, there is an important caveat: the theorem by Lyu and Li [LL20] shows convergence of gradient flow direction to a *stationary point* of the optimization in eq. (1), which might not be a global minimum; and in fact it need not even be a stationary point of the max–$\mathcal{R}$ margin problem in the function space. Moreover, the non-asymptotic behavior and rate of convergence of the learned predictor are not well-understood. In spite of these caveats, the ℓ_2 norm of the weights is an important measure of complexity and thus naturally motivates the study of $\mathcal{R}_\Phi(f)$.

Our goal in this work is to study $\mathcal{R}_\Phi(f)$ for linear functions expressed as linear convolutional networks. While we do not theoretically study the implicit biases from gradient descent variants, our findings nonetheless suggest properties about how convolution layers might influence these implicit biases. We empirically validate our findings for gradient descent in Section 6.

1.2 Notation

We typeface vectors, matrices, and tensors using bold characters, *e.g.,* $\mathbf{v}, \mathbf{x}, \boldsymbol{\theta}, \mathbf{W}, \boldsymbol{\mathcal{U}}$. We will use zero-based indexing and python style slicing notation to specify the sub-entries of an array variable, *e.g.,* for $\mathbf{Z} \in \mathbb{R}^{D_1 \times D_2}$ and $\forall_{d_1 \in [D_1]}, \forall_{d_2 \in [D_2]}$ (where $[D] = \{0, 1, \ldots, D-1\}$), the individual entries of $\mathbf{Z}$ are denoted as $\mathbf{Z}[d_1, d_2]$, the d_1^{th} row as $\mathbf{Z}[d_1, :] \in \mathbb{R}^{d_2}$, and the d_2^{th} column as $\mathbf{Z}[:, d_2] \in \mathbb{R}^{d_1}$. We also briefly mention some standard notation for complex numbers. Complex numbers are specified in the polar form as $z = |z|\,\mathrm{e}^{i\phi_z}$ with $|z| \in \mathbb{R}_+$ and $\phi_z \in [0, 2\pi)$; or in Cartesian form as $z = \mathrm{Re}(z) + i\,\mathrm{Im}(z)$, where $\mathrm{Re}(z), \mathrm{Im}(z) \in \mathbb{R}$ (ref. $i = \sqrt{-1}$ is the imaginary unit). The complex conjugate is denoted as $\overline{z} = |z|\,\mathrm{e}^{-i\phi_z}$. The standard inner product between $\mathbf{a}, \mathbf{b} \in \mathbb{C}^D$ is $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \overline{\mathbf{b}} = \sum_{d=0}^{D-1} \mathbf{a}[d]\overline{\mathbf{b}[d]}$, and analogously extends to matrices.

Unless qualified otherwise, we use $\|.\|$ to denote the standard Euclidean norm, *i.e.,* ℓ_2 norm for vectors and Frobenius norm for matrices. For arrays $\mathbf{a}, \mathbf{b}$, $\mathbf{a} \odot \mathbf{b}$ denotes entry-wise multiplication and $\mathbf{a} \propto \mathbf{b}$ implies proportionality upto positive scaling. Finally, we define the convolution operator $\star$ as it is used in the neural networks literature:[1]

Definition 1 (Circular convolution)**.** *For $\mathbf{u} \in \mathbb{R}^K$ and $\mathbf{v} \in \mathbb{R}^D$ with $K \leq D$, their D dimensional circular convolution,[2] denoted by $\mathbf{u} \star \mathbf{v}$, is a vector in $\mathbb{R}^D$ given as follows:*

$$\forall_{d \in [D]}, (\mathbf{u} \star \mathbf{v})[d] = \frac{1}{\sqrt{D}} \sum_{k=0}^{K-1} \mathbf{u}[k]\mathbf{v}[(d+k) \ mod\ D].$$

2 Multi-channel linear convolutional network

We consider two layer linear convolutional networks with multiple channels in the convolution layer. We first focus on multi-output channel convolutions with single channel inputs described below. We will discuss networks with multi-channel inputs (*e.g.,* RGB color channels) in Section 5.

The inputs to the network are vectors[3] of dimension D denoted as $\mathbf{x} \in \mathbb{R}^D$. The first layer is a convolution layer with kernel size K and number of output channel C whose weights (parameters) are denoted by $\mathbf{U} \in \mathbb{R}^{K \times C}$. The output of the convolution layer on input $\mathbf{x}$ is denoted as $h(\mathbf{U}; \mathbf{x}) \in \mathbb{R}^{D \times C}$, and is given by $h(\mathbf{U}; \mathbf{x})[:, c] = \mathbf{U}[:, c] \star \mathbf{x}$ for all $0 \leq c \leq C-1$. The second layer is a single output linear layer with weights denoted by $\mathbf{V} \in \mathbb{R}^{D \times C}$. Thus, for input $\mathbf{x}$, the output of the network $\Phi(\mathbf{U},\mathbf{V}; \mathbf{x})$ with weights $(\mathbf{U}, \mathbf{V})$ is given by:

$$\Phi(\mathbf{U},\mathbf{V};\mathbf{x}) = \langle \mathbf{V}, h(\mathbf{U};\mathbf{x}) \rangle = \sum_{c=0}^{C-1} \langle \mathbf{V}[:,c], \mathbf{U}[:,c] \star \mathbf{x} \rangle. \tag{3}$$

Since, the network described above does not have any non-linearity, the function computed by the network can be equivalently represented by $w(\mathbf{U}, \mathbf{V}) \in \mathbb{R}^D$ such that $\forall \mathbf{x}$, $\Phi(\mathbf{U},\mathbf{V}; \mathbf{x}) = \langle w(\mathbf{U}, \mathbf{V}), \mathbf{x} \rangle$. Using simple algebraic manipulations on eq. (3), one can derive $w(\mathbf{U}, \mathbf{V})$ as follows:

$$w(\mathbf{U},\mathbf{V}) = \sum_{c=0}^{C-1} \Big(\mathbf{U}[:, c] \star \mathbf{V}[:, c]^{\downarrow} \Big)^{\downarrow}, \tag{4}$$

where $\mathbf{z}^{\downarrow}$ denotes the flipped vector of $\mathbf{z} \in \mathbb{R}^D$, given by $\mathbf{z}^{\downarrow}[d] = \mathbf{z}[D-d-1]$ for $d = 0, 1, \ldots D-1$.

Remark. *Even for the smallest network in this class with $K = C = 1$, any linear predictor $\mathbf{w} \in \mathbb{R}^D$ can realized as $w(\mathbf{U}, \mathbf{V})$ in eq. (4) (e.g., using $\mathbf{U} = 1, \mathbf{V} = \mathbf{w}$). In fact, every linear predictor can be represented by multiple networks with different weights $\mathbf{U}, \mathbf{V}$.*

[1] The operator mod refers to the modulo operation for integer division, *i.e.,* p mod $D = p - D \left\lfloor \frac{p}{D} \right\rfloor$.

[2] Definition 1 corresponds to circular padding, which simplifies analysis. For convolutions with zero-padding, there will be different edge effects, but we expect qualitatively similar behavior for small padding sizes. We also use a scaling of $1/\sqrt{D}$–this is merely to simplify notation and does not change the analysis.

[3] For simplicity we consider 1D vectors $\mathbf{x} \in \mathbb{R}^D$ as inputs, but all our results can be extended to 2D inputs $\mathbf{x} \in \mathbb{R}^{W \times H}$, such as images, with the corresponding 2D convolutional operator.

Fourier domain representation. The convolution operation in Definition 1 permits a simple formulation in the Fourier domain arising from the Convolution Theorem. We consider discrete Fourier transforms (DFTs) over $\mathbb{R}^D$. Let $\mathbf{F} \in \mathbb{C}^{D\times D}$ denote the *unitary DFT matrix* for $\mathbb{R}^D$, *i.e.*, $\mathbf{F}[k,l] = \frac{1}{\sqrt{D}}\mathrm{e}^{\frac{-2\pi ikl}{D}}$ for $0 \leq k,l < D$; and for any $1 \leq K \leq D$, let $\mathbf{F}_K \in \mathbb{C}^{D\times K}$ denote the submatrix of $\mathbf{F}$ with the first K columns. For a vector $\mathbf{a} \in \mathbb{R}^K$, we denote its D dimensional Fourier representation as $\widehat{\mathbf{a}} = \mathbf{F}_K\mathbf{a} \in \mathbb{C}^D$. The Convolutional Theorem in Fourier domain implies $\mathbf{F}(\mathbf{a} \star \mathbf{b}) = \overline{\widehat{\mathbf{a}}} \odot \widehat{\mathbf{b}}$.

The discrete Fourier transform $\widehat{w}(\mathbf{U},\mathbf{V}) := \mathbf{F}w(\mathbf{U},\mathbf{V})$ of the linear predictor $w(\mathbf{U},\mathbf{V})$ realized by our network (eq. 4) can now be expressed as follows: Let $\widehat{\mathbf{U}} = \mathbf{F}_K\mathbf{U} \in \mathbb{C}^{D\times C}$ and $\widehat{\mathbf{V}} = \mathbf{F}\mathbf{V} \in \mathbb{C}^{D\times C}$ denote the D dimensional Fourier representation of $\mathbf{U},\mathbf{V}$, respectively. We have,

$$\widehat{w}(\mathbf{U},\mathbf{V}) = \sum_{c=0}^{C-1} \widehat{\mathbf{U}}[:,c] \odot \widehat{\mathbf{V}}[:,c] = \mathrm{diag}(\widehat{\mathbf{U}}\widehat{\mathbf{V}}^\top). \tag{5}$$

3 Induced regularizer in the function space

For the network Φ described above we now turn to the function space view of controlling the ℓ_2 norm of the weights $(\mathbf{U},\mathbf{V})$. We recall our discussion in Section 1.1 that this inductive bias is captured by the *induced regularizer* defined as the function space representation cost (eq. (2)). In the case of our network Φ, the function class realized is exactly the set of linear predictors in $\mathbb{R}^D$. Thus, for any $\mathbf{w} \in \mathbb{R}^D$ the induced regularizer is given as follows:

$$\mathcal{R}_{K,C}(\mathbf{w}) := \min_{\mathbf{U}\in\mathbb{R}^{K\times C},\mathbf{V}\in\mathbb{R}^{D\times C}} \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2 \quad \text{s.t.,} \quad w(\mathbf{U},\mathbf{V}) = \mathbf{w}. \tag{6}$$

Remark 1. *It immediately follows from eq.* (6) *are that $\mathcal{R}_{K,C}(\mathbf{w})$ is weakly decreasing in both K and C, i.e., $\forall_C$, $\mathcal{R}_{1,C}(\mathbf{w}) \geq \mathcal{R}_{2,C}(\mathbf{w}) \geq \ldots \mathcal{R}_{D,C}(\mathbf{w})$ and $\forall_K$, $\mathcal{R}_{K,1}(\mathbf{w}) \geq \mathcal{R}_{K,2}(\mathbf{w}) \geq \ldots$.*

Our goal is to understand this measure for different choices of K and C. Before we present our main results for multi-channel networks in Sections 4–5, we discuss some special cases which highlight how the induced regularizer changes with kernel size K in single channel networks.

Induced regularizer in special cases. Consider single channel networks ($C = 1$): the weights are $\mathbf{U} \in \mathbb{R}^K, \mathbf{V} \in \mathbb{R}^D$ and the Fourier coefficients of linear predictor $w(\mathbf{U},\mathbf{V})$ is $\widehat{w}(\mathbf{U},\mathbf{V}) = \widehat{\mathbf{U}} \odot \widehat{\mathbf{V}}$.

For any kernel size K, we can obtain a lower bound on the induced regularizer as $\mathcal{R}_{K,1}(\mathbf{w}) \geq 2\|\widehat{\mathbf{w}}\|_1$ using $\forall_{d\in[D]}, |\widehat{\mathbf{U}}[d]|^2 + |\widehat{\mathbf{V}}[d]|^2 \geq 2|\widehat{\mathbf{U}}[d] \cdot \widehat{\mathbf{V}}[d]|$. In the case of a full dimensional convolution layer with $K = D$, Gunasekar et al. [GLSS18b] showed that this lower bound is indeed tight:

Lemma 1 ($K = D$)**.** *[Lemma 7 in GLSS18b] For any $\mathbf{w} \in \mathbb{R}^D$, $\mathcal{R}_{D,1}(\mathbf{w}) = 2\|\widehat{\mathbf{w}}\|_1$.*

The proof of Lemma 1 uses the fact that for $K = D$, the weights $\mathbf{U},\mathbf{V} \in \mathbb{R}^D$ are unconstrained and can be chosen such that $\forall_{d\in D}$, $|\widehat{\mathbf{U}}[d]| = |\widehat{\mathbf{V}}[d]| = \sqrt{|\widehat{\mathbf{w}}[d]|}$ leading to the optimal cost of $2\|\widehat{\mathbf{w}}\|_1$.

For networks with smaller kernels, the argument breaks as $\widehat{\mathbf{U}} = \mathbf{F}_K\mathbf{U}$ is constrained to be in a $K < D$ dimensional space spanned by the columns of $\mathbf{F}_K$. Thus, it is not always possible to choose weights satisfying $|\widehat{\mathbf{U}}[d]| = \sqrt{|\widehat{\mathbf{w}}[d]|}$. It is easiest to see this is case of kernel size $K = 1$, where $\mathbf{U} \in \mathbb{R}^1$ is a scalar. Since $\widehat{\mathbf{U}} \propto [1, 1, \ldots, 1]$, the constraints imply that $\widehat{\mathbf{V}} \propto \widehat{\mathbf{w}}$. By choosing u_0 optimally, we can show the following lemma (full proof is provided in the Appendix B):

Lemma 2 ($K = 1$)**.** *For any $\mathbf{w} \in \mathbb{R}^D$, it holds that $\mathcal{R}_{1,1}(\mathbf{w}) = 2\sqrt{D}\|\widehat{\mathbf{w}}\|_2 = 2\sqrt{D}\|\mathbf{w}\|_2$.*

The induced regularizer thus behaves fundamentally differently for $K = D$ and $K = 1$. In particular, the ℓ_2 regularization of $\mathcal{R}_{1,1}(\mathbf{w})$ does not induce sparse solutions, while the ℓ_1 regularization of $\mathcal{R}_{D,1}(\mathbf{w})$ promotes sparsity in the Fourier basis. We show in Section 4 that for general K the induced regularizer is a norm that interpolates between ℓ_2 and ℓ_1 norms in Fourier space.

Since $K = 1$ and $K = D$ permit closed-form solutions in the Fourier space, one might hope to obtain similarly clean characterizations for other kernel sizes as well. Unfortunately, even for $K = 2$, we show that $\mathcal{R}_{2,1}(\mathbf{w})$ takes a much more complex form. In the following lemma we show that for $K = 2$ the characterization of $\mathcal{R}_{2,1}(\mathbf{w})$ in Fourier space involves a maximization over a high-degree rational function, and is thus unlikely to admit clean closed-form solutions.

Lemma 3. *For any $\mathbf{w} \in \mathbb{R}^D$, it holds that:*

$$\mathcal{R}_{2,1}(\mathbf{w}) = 2\sqrt{D}\sqrt{\inf_{\alpha\in(-1,1)} \sum_{d=0}^{D-1} \frac{|\widehat{\mathbf{w}}[d]|^2}{1+\alpha\cos(2\pi d/D)}}.$$

Although Lemma 3 does not yield closed form solutions for $\mathcal{R}_{2,1}(\mathbf{w})$, we observe that it hints at some form of band-pass frequency structure: for any α from the inner optimization, the resulting regularizer is a weighted sum of Fourier coefficients such that the nearby frequency components of $\widehat{\mathbf{w}}$ are weighted with nearby values. This band-pass nature was also observed in a complementary result by Yun et al. [YKM21] in the context of implicit bias from gradient descent on a single data point: for any $\mathbf{x}$, it was shown that $\min_{\mathbf{w}} \mathcal{R}_{2,1}(\mathbf{w})$ s.t., $\mathbf{w}^\top\mathbf{x} > 1$ corresponds to a low-pass or high pass filter depending on the sign of $\mathbf{x}^\top\mathbf{x}^{\downarrow}$.

Even though we do not obtain closed form solutions for all kernel sizes K, we derive important properties about the induced regularizer for general kernel sizes in the following sections.

4 Main technical tool: SDP formulation of induced regularizer

To investigate the induced regularizer for general kernel sizes, we construct a semidefinite program (SDP) relaxation that turns out to be independent of C. In this section, we describe and analyze this SDP formulation for networks on inputs with a single channel. (We discuss generalizations to the case of multi-channel inputs in Section 5.)

We first reformulate $\mathcal{R}_{K,C}$ as an SDP with a rank constraint, which immediately motivates an SDP relaxation that provides a lower bound on $\mathcal{R}_{K,C}$. As we will show in Theorem 4, this SDP relaxation is actually tight for all K and C, which enables us to deduce a number of interesting properties of the induced regularizer.

$\mathcal{R}_{K,C}$ as an SDP with a rank constraint. Combining the definition of $\mathcal{R}_{K,C}(\mathbf{w})$ in eq. (6) with the Fourier representation of $w(\mathbf{U},\mathbf{V})$ in eq. (5), we have the following:

$$\mathcal{R}_{K,C}(\mathbf{w}) = \min_{\mathbf{U}\in\mathbb{R}^{K\times C},\mathbf{V}\in\mathbb{R}^{D\times C}} \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2 \quad \text{s.t.,} \quad \operatorname{diag}(\widehat{\mathbf{U}}\widehat{\mathbf{V}}^\top) = \widehat{\mathbf{w}}. \tag{7}$$

We observe that we can express the objective of eq. (7) as $\|\mathbf{U}\|^2 + \|\mathbf{V}\|^2 = \langle \mathbf{U}\mathbf{U}^\top, \mathbf{I}\rangle + \langle \mathbf{V}\mathbf{V}^\top, \mathbf{I}\rangle$. Similarly, the constraints are give by $\forall_{d\in[D]}$, $\langle \widehat{\mathbf{U}}\widehat{\mathbf{V}}^\top, \mathbf{e}_d\mathbf{e}_d^\top\rangle = \widehat{\mathbf{w}}[d]$, where $\{\mathbf{e}_d\}_{d\in[D]}$ denotes the standard basis, or alternatively, as $\forall_{d\in[D]}, \langle \mathbf{U}\mathbf{V}^\top, \mathbf{Q}_d\rangle = \widehat{\mathbf{w}}[d]$, where $\mathbf{Q}_d := \overline{\mathbf{F}}_K^\top \mathbf{e}_d\mathbf{e}_d^\top \overline{\mathbf{F}} \in \mathbb{C}^{K\times D}$.

The optimization in eq. (7) over $\mathbf{U} \in \mathbb{R}^{K\times C}, \mathbf{V} \in \mathbb{R}^{D\times C}$, can thus be specified in terms of a rank C positive semi-definite matrix $\mathbf{Z} \in \mathbb{R}^{(D+K)\times(D+K)}$ that we define below:

$$\mathbf{Z} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}\begin{bmatrix}\mathbf{U}^\top & \mathbf{V}^\top\end{bmatrix} = \begin{bmatrix}\mathbf{U}\mathbf{U}^\top & \mathbf{U}\mathbf{V}^\top\\ \mathbf{V}\mathbf{U}^\top & \mathbf{V}\mathbf{V}^\top\end{bmatrix} \succcurlyeq 0. \tag{8}$$

The objective and constraints of eq. (7) can now be expressed as linear functions of $\mathbf{Z}$ as $\langle \mathbf{Z}, \mathbf{I} \rangle$, and $\forall_{d\in[D]}$, $\langle \mathbf{Z}, \mathbf{A}_d^{\text{real}} \rangle = 2 \cdot \text{Re}(\widehat{\mathbf{w}}[d])$ and $\langle \mathbf{Z}, \mathbf{A}_d^{\text{img}} \rangle = 2 \cdot \text{Im}(\widehat{\mathbf{w}}[d])$, respectively, where we define $(\mathbf{A}_d^{\text{real}}, \mathbf{A}_d^{\text{img}})$ as follows:

$$\mathbf{A}_d^{\text{real}} = \begin{bmatrix} 0_K & \mathbf{Q}_d \\ \overline{\mathbf{Q}}_d^\top & 0_D \end{bmatrix} \text{ and } \mathbf{A}_d^{\text{img}} = \begin{bmatrix} 0_K & i \cdot \mathbf{Q}_d \\ -i \cdot \overline{\mathbf{Q}}_d^\top & 0_D \end{bmatrix}.$$

Now, we can formulate $\mathcal{R}_{K,C}(\mathbf{w})$ as follows:

$$\begin{aligned} \mathcal{R}_{K,C}(\mathbf{w}) \;=\; \min_{\mathbf{Z}\succcurlyeq 0} \langle \mathbf{Z}, \mathbf{I} \rangle \quad \text{s.t.,} \quad & \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\text{real}} \rangle = 2\,\text{Re}(\widehat{\mathbf{w}}[d]) \\ & \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\text{img}} \rangle = 2\,\text{Im}(\widehat{\mathbf{w}}[d]) \\ & \text{rank}(\mathbf{Z}) \le C. \end{aligned} \tag{9}$$

The formulation in eq. (9) is non-convex due to the rank constraint. We obtain a natural convex relaxation by dropping the rank constraint, leading to the following SDP:

$$\begin{aligned} \mathcal{R}_K^{\text{SDP}}(\mathbf{w}) \;=\; \min_{\mathbf{Z}\succcurlyeq 0} \langle \mathbf{Z}, \mathbf{I} \rangle \quad \text{s.t.,} \quad & \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\text{real}} \rangle = 2\,\text{Re}(\widehat{\mathbf{w}}[d]) \\ & \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\text{img}} \rangle = 2\,\text{Im}(\widehat{\mathbf{w}}[d]). \end{aligned} \tag{10}$$

Remark. *By construction, the relaxation provides lower bounds on the induced regularizer: for any $K \le D$, any C, and any $\mathbf{w} \in \mathbb{R}^D$, it holds that $\mathcal{R}_{K,C}(\mathbf{w}) \ge \mathcal{R}_K^{SDP}(\mathbf{w})$.*

Remark. *The symmetry properties of Fourier coefficients of real signals gives us that for any D and any $\mathbf{w} \in \mathbb{R}^D$, $\widehat{\mathbf{w}}[p] = \overline{\widehat{\mathbf{w}}[D-p]}$ for $p \in [D]$. Thus, although the optimization problems in* (9) *and* (10) *are specified with $2 \cdot D$ constraints for simplicity, only D of them are unique.*

4.1 Tightness of the SDP Relaxation

Our main technical result is that for any kernel size K, the SDP relaxation is tight (in the case of networks with single-channel inputs). Thus, the induced regularizer $\mathcal{R}_{K,C}$ is equivalent to an SDP that only depends upon on the kernel size K.

Theorem 4. *For any $K \le D$, any C, and any $\mathbf{w} \in \mathbb{R}^D$, it holds that $\mathcal{R}_{K,C}(\mathbf{w}) = \mathcal{R}_K^{SDP}(\mathbf{w})$.*

Proof sketch. We can show directly from the KKT conditions that *any* minimizer $\mathbf{Z}$ of the SDP must have rank at most K. However, to prove Theorem 4, we need to show that there exists a rank 1 solution that has the same objective value as $\mathbf{Z}$ and satisfies the SDP constraints—this does not follow directly from the KKT conditions. Constructing this rank 1 solution is the main contribution in the proof of Theorem 4. In particular, the following lemma is a key intermediate result about the convolutional operation and is of independent interest beyond this paper.

Lemma 5. *For any $1 \le K \le D$, and for any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$, there exists a vector $\mathbf{c} \in \mathbb{R}^K$ such that $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b} = \mathbf{c} \star \mathbf{c}$, where convolutions are w.r.t. dimension D.*

For $K = D$, Lemma 5 follows easily from the Fourier space representation (using $\mathbf{z} \star \mathbf{z} = |\widehat{\mathbf{z}}|^2$), since in this case $\widehat{\mathbf{c}}$ is unconstrained and can be explicitly constructed as the square root of the Fourier transform of $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b}$. However, this construction does not generalize to kernel sizes $K < D$. In fact, $\mathbf{c}$ does not appear to have an explicit closed-form for general $K < D$.

In our proof, we show existence of $\mathbf{c}$ by leveraging the polynomial representation of convolutions. Using the polynomial representations, Lemma 5 can be written in the following form: for any real-coefficient polynomials $p_{\mathbf{a}}, p_{\mathbf{b}}$ with degree at most $K-1$, there exists a real-coefficient polynomial $p_{\mathbf{c}}$ of degree at most $K-1$ such that:

$$x^{K-1}p_{\mathbf{c}}(x)p_{\mathbf{c}}(1/x) = x^{K-1}p_{\mathbf{a}}(x)p_{\mathbf{a}}(1/x) + x^{K-1}p_{\mathbf{b}}(x)p_{\mathbf{b}}(1/x).$$

The remainder of the proof involves implicitly constructing $p_{\mathbf{c}}$ in terms of its roots (and leading coefficient). To do so, we show that the roots of the polynomial $x^{K-1}p_{\mathbf{a}}(x)p_{\mathbf{a}}(1/x) + x^{K-1}p_{\mathbf{b}}(x)p_{\mathbf{b}}(1/x)$ satisfy certain structural properties which allow us to establish the existence of the desired real-coefficient polynomial $p_{\mathbf{c}}$. The full proof of Theorem 4 can be found in the Appendix C. □

Consequences of Theorem 4. The first consequence of Theorem 4 is that although the optimization in eq. (6) is non-convex, the SDP formulation allows us to efficiently compute $\mathcal{R}_{K,C}(\mathbf{w})$ exactly. In the remainder of the section, we discuss a number of other interesting properties that we can deduce from Theorem 4. Proofs of the following results are in Appendix D.

4.1.1 $\mathcal{R}_{K,C}$ is independent of number of output channels C

Theorem 4 directly implies that $\mathcal{R}_{K,C}$ is independent of C. This means that the linear predictors obtained by fitting training data and minimizing $\mathcal{R}_{K,C}(\mathbf{w})$ will be invariant to C (apart from $\mathcal{R}_{K,C}(\mathbf{w})$ having multiple minimizers). This has potential implications for implicit bias from gradient descent. Recall the discussion in Section 1.1 that (subject to the caveats mentioned earlier) *suggests* that implicit bias from gradient descent is related to the minimizing $\mathcal{R}_{K,C}(\mathbf{w})$. Based on our result, we can hypothesize that that for linear networks with single channel input, the number output channels does not influence the asymptotic predictor learned from gradient descent. In Section 6 we show empirical evidence supporting this hypothesis.

4.1.2 $\mathcal{R}_{K,C}$ is a norm

Another interesting corollary of Theorem 4 is that the induced regularizer is a norm for any K.

Corollary 6. *For $K \leq D$ and any C, $\mathcal{R}_{K,C}(\mathbf{w})$ is a norm.*

For the end cases of $K = 1$ and $K = D$, this norm can be explicitly specified: $\mathcal{R}_{1,C}(\mathbf{w}) = 2\sqrt{D}\|\widehat{\mathbf{w}}\|$ (Lemma 2) and $\mathcal{R}_{D,C}(\mathbf{w}) = 2\|\widehat{\mathbf{w}}\|_1$ (Lemma 1), respectively. For intermediate kernel sizes, $\mathcal{R}_{K,C}(\mathbf{w})$ is a norm that interpolates between the ℓ_2 norm and the ℓ_1 norm of the Fourier coefficients of the linear predictor.

We further use the SDP in (10) to compute upper and lower bounds on $\mathcal{R}_{K,C}(\mathbf{w})$ in terms of the ℓ_2 norm and ℓ_1 in Fourier space.

Lemma 7. *For any $K \leq D$, any C, and any $\mathbf{w} \in \mathbb{R}^D$:*

$$2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_2 \leq \mathcal{R}_{K,C}(\mathbf{w}) \leq 2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$$

$$2\|\widehat{\mathbf{w}}\|_1 \leq \mathcal{R}_{K,C}(\mathbf{w}) \leq 2\sqrt{\left\lceil\frac{D}{K}\right\rceil}\|\widehat{\mathbf{w}}\|_1.$$

Remark. *For the lower bounds,* $2\|\widehat{\mathbf{w}}\|_1$ *is tight for* $\mathbf{w} = [1, 0, \ldots, 0]$ *and* $2\sqrt{\frac{D}{K}}\|\mathbf{w}\|$ *is tight for* $\mathbf{w} = [1, 1, \ldots, 1]$. *For the upper bounds,* $2\sqrt{\lceil \frac{D}{K} \rceil}\|\widehat{\mathbf{w}}\|_1$ *is tight when* $K \mid D$ *for patterned vectors (see Lemma 8), and* $2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$ *is tight for* $[1, 0, \ldots, 0]$.

Lemma 7 demonstrates that when K is a small constant, $\mathcal{R}_{K,C}(\mathbf{w})$ is multiplicatively close to $\mathcal{R}_{1,C}(\mathbf{w}) = 2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$. On the other hand, once K is comparable to D, $\mathcal{R}_{K,C}(\mathbf{w})$ is within a constant factor of $\mathcal{R}_{D,C}(\mathbf{w}) = 2\|\widehat{\mathbf{w}}\|_1$.

4.1.3 $\mathcal{R}_{K,C}(\mathbf{w})$ for patterned vectors

Aside from general bounds on $\mathcal{R}_{K,C}$, the SDP formulation can also be used to analyze the behavior of the induced regularizer of special classes of vectors. One interesting case is of *patterned vectors* described as follows: Consider vectors of the form $\mathbf{w}(\mathbf{p}) = [\mathbf{p}, \mathbf{p}, \ldots, \mathbf{p}] \in \mathbb{R}^D$ consisting of repetitions of a P dimensional pattern $\mathbf{p} \in \mathbb{R}^P$. A useful property of linear predictors of this form is that they incorporate invariance to periodic translations.

We show an interesting relation between the representation cost $\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p}))$ of realizing patterned vectors in $\mathbb{R}^D$ and the analogous cost (denoted as $\mathcal{R}^{(P)}_{K,C}(\mathbf{p})$) of realizing $\mathbf{p}$ as a linear predictor in $\mathbb{R}^P$ using a network with the same values of K and C.

Lemma 8. *Consider vectors* $\mathbf{w}(\mathbf{p}) = [\mathbf{p}, \mathbf{p}, \ldots, \mathbf{p}] \in \mathbb{R}^D$ *specified by* $\mathbf{p} \in \mathbb{R}^P$ *s.t.,* P *divides* D.
(a) For any $K \leq P$*, it holds that* $\forall C$*:*

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) = \frac{D}{P} \cdot \mathcal{R}^{(P)}_{K,1}(\mathbf{p}).$$

(b) For $P \leq K \leq D$ *if* $K = P \cdot T$ *for integer* T*, then* $\forall C$*:*

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) = 2\frac{D}{\sqrt{T}P}\|\widehat{\mathbf{p}}\|_1 = 2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_1.$$

We see that the induced regularizer of repeated patterned vectors is closely related to that of the pattern itself. In particular, for $K \leq P$, we have $\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) \propto \mathcal{R}^{(P)}_{K,1}(\mathbf{p})$.

5 Networks with multi-channel inputs

While we focused on networks with single-channel inputs in the previous sections, we now expand our results to networks with multiple input channels (*e.g.,* RGB color channels). How do these conclusions change for multiple input channels? How does the induced regularizer, now denoted as $\mathcal{R}_{K,C,R}$, depend on the number input channels R?

We again consider two layer convolutional networks akin to Section 2. We first introduce additional notation: The multi-channel inputs are denotes as $\mathbf{X} \in \mathbb{R}^{D \times R}$, where R denotes the number of input channels. The convolutional first layer now has kernel size K, output channel size C and input channel size R with weights denoted by a set of R matrices $\boldsymbol{\mathcal{U}} = \{\mathbf{U}_r\}_{r \in [R]}$ with $\mathbf{U}_r \in \mathbb{R}^{K \times C}$. The output of this convolution layer $h(\boldsymbol{\mathcal{U}}; \mathbf{X}) \in \mathbb{R}^{D \times C}$ is given as follows:

$$\forall_{c \in C}, \; h(\boldsymbol{\mathcal{U}}; \mathbf{X})[:, c] = \sum_{r=0}^{R-1} \mathbf{U}_r[:, c] \star \mathbf{X}[:, r]. \tag{11}$$

The second layer is the same as before: a single output linear layer with weights $\mathbf{V} \in \mathbb{R}^{D\times C}$. We denote the equivalent linear predictor for this network by $W(\boldsymbol{\mathcal{U}}, \mathbf{V}) \in \mathbb{R}^{D\times R}$. Following similar calculations as for single input channels, $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ in signal and Fourier domain (denoted as $\widehat{W}(\boldsymbol{\mathcal{U}}, \mathbf{V}) = \mathbf{F}W(\boldsymbol{\mathcal{U}}, \mathbf{V})$) are given as follows:

$$\forall_{r\in[R]},\ W(\boldsymbol{\mathcal{U}}, \mathbf{V})[:, r] = \sum_{c=0}^{C-1} \Big(\mathbf{U}_r[:, c] \star \mathbf{V}[:, c]^{\downarrow}\Big)^{\downarrow}, \text{ and } \widehat{W}(\boldsymbol{\mathcal{U}}, \mathbf{V})[:, r] = \mathrm{diag}(\widehat{\mathbf{U}}_r \widehat{\mathbf{V}}^\top). \tag{12}$$

For multi-channel inputs, the set of all linear predictors is the space of matrices $\mathbf{W} \in \mathbb{R}^{D\times R}$, and we define the induced complexity measure over this matrix space as follows:

$$\mathcal{R}_{K,C,R}(\mathbf{W}) := \inf_{\boldsymbol{\mathcal{U}},\mathbf{V}} \sum_{r\in[R]} \|\mathbf{U}_r\|^2 + \|\mathbf{V}\|^2 \quad \text{s.t.,} \quad W(\boldsymbol{\mathcal{U}}, \mathbf{V}) = \mathbf{W}. \tag{13}$$

5.1 Role of output channel size C

For multi-channel inputs, we first observe that multiple output channels can be necessary to realize all linear maps. To see this, we show that the sub-network corresponding to each output channel can realize a matrix in $\mathbb{R}^{D\times R}$ of rank at most K, which places an upper bound on the total rank achievable by the full network (see a proof in Appendix E.1). This implies the following lemma:

Lemma 9. *For any K, C and R, in order for the the network represented by $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ in eq.* (12) *to realizes all linear maps in $\mathbb{R}^{D\times R}$ it is necessary that $K \cdot C \geq \min\{R, D\}$.*

In contrast to single input channels, Lemma 9 demonstrates that, the model class realized by linear convolutional networks over multi-channel inputs, and consequently the induced regularizer, *does* depend on number of output channels C. Nonetheless, similar to single input channel networks, we can again obtain an SDP relaxation $\mathcal{R}^{\mathrm{SDP}}_{K,R}(\mathbf{W})$ for $\mathcal{R}_{K,C,R}(\mathbf{W})$ that is independent of C.

5.2 SDP relaxation for multi-channel input networks

The SDP relaxation for $\mathcal{R}_{K,C,R}$ is derived similarly to for networks with a single input channel. For any $\boldsymbol{\mathcal{U}} = \{\mathbf{U}_r \in \mathbb{R}^{K\times C}\}_{r\in[R]}, \mathbf{V} \in \mathbb{R}^{D\times C}$, we define a rank C positive semidefinite matrix $\mathbf{Z} \in \mathbb{R}^{(D+K\cdot R)\times(D+K\cdot R)}$, that represents:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{U}_0 \\ \mathbf{U}_1 \\ \dots \\ \mathbf{U}_R \\ \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{U}_0^\top & \mathbf{U}_1^\top & \dots & \mathbf{U}_R^\top & \mathbf{V}^\top \end{bmatrix} \succcurlyeq 0.$$

We also define Hermitian matrices $\mathbf{A}_{d,r}^{\text{real}}, \mathbf{A}_{d,r}^{\text{img}} \in \mathbb{R}^{(D+K\cdot R)\times(D+K\cdot R)}$ for $d \in [D], r \in [R]$. If we let $\mathbf{Q}_d = \overline{\mathbf{F}}_K^\top \mathrm{e}_d \mathrm{e}_d^\top \overline{\mathbf{F}}$, these matrices take the following form:

$$\mathbf{A}_{d,0}^{\text{real}} = \left[\begin{array}{cccc|c} & & & & \mathbf{Q}_d \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & & \mathbf{0} \\ & & & & \vdots \\ & & & & \mathbf{0} \\ \hline \overline{\mathbf{Q}}_d & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0}_{D\times D} \end{array}\right], \quad \mathbf{A}_{d,0}^{\text{img}} = \left[\begin{array}{cccc|c} & & & & i\cdot\mathbf{Q}_d \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & & \mathbf{0} \\ & & & & \vdots \\ & & & & \mathbf{0} \\ \hline -i\cdot\overline{\mathbf{Q}}_d & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0}_{D\times D} \end{array}\right],$$

$$\mathbf{A}_{d,1}^{\text{real}} = \left[\begin{array}{cccc|c} & & & & \mathbf{0} \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & & \mathbf{Q}_d \\ & & & & \vdots \\ & & & & \mathbf{0} \\ \hline \mathbf{0} & \overline{\mathbf{Q}}_d & \dots & \mathbf{0} & \mathbf{0}_{D\times D} \end{array}\right], \quad \mathbf{A}_{d,1}^{\text{img}} = \left[\begin{array}{cccc|c} & & & & \mathbf{0} \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & & i\cdot\mathbf{Q}_d \\ \mathbf{0} & & & & \\ & & & & \vdots \\ & & & & \mathbf{0} \\ \hline \mathbf{0} & -i\cdot\overline{\mathbf{Q}}_d & \dots & \mathbf{0} & \mathbf{0}_{D\times D} \end{array}\right],$$

$$\vdots \qquad\qquad\qquad\qquad \vdots$$

$$\mathbf{A}_{d,R}^{\text{real}} = \left[\begin{array}{cccc|c} & & & & \mathbf{0} \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & & \mathbf{0} \\ & & & & \vdots \\ & & & & \mathbf{Q}_d \\ \hline \mathbf{0} & \mathbf{0} & \dots & \overline{\mathbf{Q}}_d & \mathbf{0}_{D\times D} \end{array}\right], \quad \mathbf{A}_{d,R}^{\text{img}} = \left[\begin{array}{cccc|c} & & & & \mathbf{0} \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & & \mathbf{0} \\ & & & & \vdots \\ & & & & i\cdot\mathbf{Q}_d \\ \hline \mathbf{0} & \mathbf{0} & \dots & -i\cdot\overline{\mathbf{Q}}_d & \mathbf{0}_{D\times D} \end{array}\right].$$

We also provide a more formal description of these matrices. We provide a block-wise description of only the upper diagonal blocks, with lower diagonal blocks filled to satisfy the Hermitian matrix property. Additionally, for matrices $\{\mathbf{A}_{d,r}^{\text{real}}, \mathbf{A}_{d,r}^{\text{img}}\}_{d\in[D],r\in[R]}$ any unspecified block is by default treated as zero matrix $\mathbf{0}$ of appropriate dimension:

- For $r_1, r_2 \in [R]$ with $r_2 \geq r_1$, the $K \times K$ block with indices $(r_1 : (r_1+1)K)$ along rows and $(r_2 : (r_2+1)K)$ along columns is given as

$$\mathbf{Z}[r_1 : (r_1+1)K, r_2 : (r_2+1)K] = \mathbf{U}_{r_1}\mathbf{U}_{r_2}^\top;$$

- For $r \in [R]$, the $K \times D$ blocks with indices $(r : (r+1)K)$ along rows and $(RK : (D+RK))$ along column, are given as follows:

$$\mathbf{Z}[r : (r+1)K, RK : (D+RK)] = \mathbf{U}_r\mathbf{V}^\top.$$

  Further,

$$\forall_{d\in[D]},\ \mathbf{A}_{d,r}^{\text{real}}[r : (r+1)K, RK : (D+RK)] = \mathbf{Q}_d, \text{ and}$$
$$\mathbf{A}_{d,r}^{\text{img}}[r : (r+1)K, RK : (D+RK)] = i\cdot\mathbf{Q}_d.$$

  Note that for $r' \neq r$, the corresponding blocks in $\mathbf{A}_{d,r'}^{\text{real}}, \mathbf{A}_{d,r'}^{\text{img}}$ remain the default zero.
- Finally, the lower-right $D \times D$ block is given as

$$\mathbf{Z}[RK : (D+RK), RK : (D+RK)] = \mathbf{V}\mathbf{V}^\top.$$

Using this notation, we consider the following SDP relaxation of $\mathcal{R}_{K,C,R}(\mathbf{W})$ in terms of Fourier coefficients $\widehat{\mathbf{W}} = \mathbf{F}\mathbf{W}$:

$$\begin{aligned}\mathcal{R}^{\mathrm{SDP}}_{K,R}(\mathbf{W}) \;=\; \min_{\mathbf{Z}\succcurlyeq 0} \quad & \langle \mathbf{Z}, \mathbf{I}\rangle \\ \text{s.t.} \quad & \forall d\in[D], r\in[R] \;\; \langle \mathbf{Z}, \mathbf{A}^{\mathrm{real}}_{d,r}\rangle = 2\mathrm{Re}(\widehat{\mathbf{W}}[d,r]) \\ & \forall d\in[D], r\in[R] \;\; \langle \mathbf{Z}, \mathbf{A}^{\mathrm{img}}_{d,r}\rangle = 2\mathrm{Im}(\widehat{\mathbf{W}}[d,r]).\end{aligned} \tag{14}$$

We can check that the SDP formulation with an additional rank constraint of $\mathrm{rank}\,(\mathbf{Z}) \leq C$ is equivalent $\mathcal{R}_{K,C,R}(\mathbf{W})$ and the SDP thus provides a lower bound: *i.e.,* $\forall_{\mathbf{W}},\ \mathcal{R}_{K,C,R}(\mathbf{W}) \geq \mathcal{R}^{\mathrm{SDP}}_{K,R}(\mathbf{W})$.

5.3 Tightness of SDP Relaxation

Unlike networks with single channel input, the SDP relaxation here is not always tight when $R > 1$, since sufficiently large C is required to merely realize all matrix-valued linear function over the input space. We can however show a weaker form SDP tightness from the KKT conditions when there are sufficiently many output channels:

Lemma 10. *For any $\mathbf{W} \in \mathbb{R}^{D\times R}$, and any $C \geq RK$, it holds that $\mathcal{R}_{K,C,R}(\mathbf{W}) = \mathcal{R}^{SDP}_{K,R}(\mathbf{W})$.*

Note that the above bound on C for SDP tightness is not sharp, as we showed for $R = 1$ in Theorem 4. Based on our insights from the proof of single-channel SDP tightness in Theorem 4 and additional empirical evidence in Appendix A, we conjecture that SDP tightness holds when $C \geq R$:

Conjecture 11. *For any $\mathbf{W} \in \mathbb{R}^{D\times R}$, and any $C \geq R$, it holds that $\mathcal{R}_{K,C,R}(\mathbf{W}) = \mathcal{R}^{SDP}_{K,R}(\mathbf{W})$.*

In the next subsection, we prove Conjecture 11 in the special cases of $K = 1$ and $K = D$. As a consequence, we show that once C is large enough to realize all linear maps, $\mathcal{R}_{K,C,R}(\mathbf{W})$ can be expressed as interesting closed form norms independent of C in these special cases.

5.4 Induced regularizer when $K = 1$ and $K = D$

Theorem 12. *For any $\mathbf{W} \in \mathbb{R}^{D\times R}$, and any $C \geq \min\{R, D\}$, the induced regularizer for $K = 1$ is given by the scaled nuclear norm $\|.\|_*$:*

$$\mathcal{R}_{1,C,R}(\mathbf{W}) = 2\sqrt{D}\|\mathbf{W}\|_* = 2\sqrt{D}\|\widehat{\mathbf{W}}\|_*.$$

Theorem 13. *For any $\mathbf{W} \in \mathbb{R}^{D\times R}$, and any $C \geq 1$, the induced regularizer for $K = D$ is given as follows*

$$\mathcal{R}_{D,C,R}(\mathbf{W}) = 2\|\widehat{\mathbf{W}}\|_{2,1} := \sum_{d=0}^{D-1}\sqrt{\sum_{r=0}^{R-1}\left|\widehat{\mathbf{W}}[d,r]\right|^2}.$$

From Theorems 12-13 it is evident that the number of input channels R fundamentally changes the nature of induced complexity measure in the function space and introduces additional structures along the input channels. Even in the simplest setting of scalar convolution kernels with $K = 1$, the induced regularizer is no longer a Euclidean or RKHS norm, and is instead a richer nuclear norm that encourages low-rank properties. For the case of $K = D$, the induced regularizer is group-sparse norm on the Fourier coefficients that encourages similar weighting across channels, while promoting sparsity across frequency components. In comparison to the ℓ_1 norm of all Fourier coefficients, this group-sparse norm is a more structured inductive bias for multi-channel inputs. Additionally, like with the single input channel case, we also observe that the induced bias has a more intuitive and interesting interpretation in Fourier domain which is not directly observed in the signal domain.

6 Experiments

We have thus far focused on analyzing the induced regularizer as a complexity measure arising from explicitly minimizing ℓ_2 norm of weights. While our results suggest connections to implicit bias from gradient descent on separable classification tasks as discussed in Section 1.1, this connection is subject to the caveats discussed therein. If we overlook the caveats, our findings would then have important implications for predictors learned from gradient descent. We formalize one of these implications—regarding the impact of the number of output channels—as a testable hypothesis.

Hypothesis 1. *For a separable binary classification task with R input channels, let $\mathbf{w}_{GD}$ be the predictor learned using stochastic gradient descent on a two-layer convolutional network with kernel size K, C output channels, and R input channels (where $\mathbf{w}_{GD}$ is normalized to have unit margin on the training data). Then, as long as $C \geq R$, the induced regularizer $\mathcal{R}_{K,C,R}(\mathbf{w}_{GD})$ is invariant in the number of output channels C.*

The primary goal of our experiments is to test this hypothesis. We show experimental support for the hypothesis on small linearly separable subsets of MNIST (with 128 images of size 28×28 balanced across 2 classes) and CIFAR-10 (with 512 images of size 32×32 balanced across 2 classes) datasets. Most of our experiments are for multi-channel linear convolutional networks trained using stochastic gradient descent. We also provide a some experiments on ReLU networks, where we see support of our hypothesis well beyond our theoretical study.

Throughout the experiments sections, since we cannot always compute $\mathcal{R}_{K,C,R}(\mathbf{w}_{\mathrm{GD}})$, we approximate it using the weight norms of the trained network $\widehat{\mathcal{R}}_{K,C,R}(\mathbf{w}_{\mathrm{GD}}) = \sum_{\boldsymbol{\mathcal{U}},\mathbf{V}} \|\boldsymbol{\mathcal{U}}\|^2 + \|\mathbf{V}\|^2$, where $\boldsymbol{\mathcal{U}}, \mathbf{V}$ here denote the weights of the trained network.[4] The experiments are deferred to Appendix A and we summarize the findings below.

1. *Single input channel binary classification on MNIST.* In linear networks, we compare the predictors learned by gradient descent for $K \in \{1, 3, 8, 16, 28\}$ and $C \in \{1, 2, 4, 8\}$ across 10 runs with random initialization. We see that both the values of estimated regularizer $\widehat{\mathcal{R}}$ as well as the visualization of linear predictors in signal and frequency domain are nearly invariant to C (the values overlap within one standard deviation across runs).
2. 3*–input channel binary classification on CIFAR-10.* In a similar setup to MNIST, we compare $K \in \{1, 3, 8, 20\}$ and $C \in \{1, 2, 3, 4, 8\}$. As expected from our theory, we see differences in the induced regularizer $\widehat{\mathcal{R}}$ for $C < 3$, but observe invariance to C once $C \geq 3$.
3. *ReLU networks for binary classification on MNIST.* Although our theory is only for linear networks, our hypothesis as stated above can also be tested on networks with non-linearity. We repeat our MNIST experiments on networks with ReLU non-linearity (with and without bias parameters). Interestingly, we observe that the estimated induced regularizer $\widehat{\mathcal{R}}$ is again invariant to C suggesting a broader scope for our hypothesis. Altogether, these findings support Hypothesis 1, including in networks beyond the scope of our theoretical results.
4. Apart from investigating the role of the number of output channels, we also provide support for our theoretical findings about the role of kernel size. On both MNIST and CIFAR-10, we show that larger kernel sizes favor sparsity in the frequency domain and that the learned predictors experience similarity across input channels for smaller kernel sizes.

[4]In theory $\widehat{\mathcal{R}}$ only provides an upper bound on $\mathcal{R}$—but in case of predictors learned by SGD, upon checking instances where $\mathcal{R}$ has a closed form solution, we found that the approximation is quite accurate.

7 Discussion and Related Work

Towards understanding neural networks, the representation cost or the induced regularizer viewpoint provides an abstraction to separate capacity control in the parameter space from the function space implications of the resulting inductive bias. In this paper, we showed that when minimizing ℓ_2 norm of weights, the two basic architectural components of convolutional networks—number of output channels (width) and kernel size—have interesting effects even in the simple case of two-layer linear networks. Our results inspire a broader hypothesis about the impact of number of output channels which we test and provide support for in our experiments. Our experiment show promise for applicability of our hypothesis beyond our theoretical finding, including to non-linear networks. An immediate direction for future work thus would be to expand on our experimental findings and conduct an in-depth empirical investigation of the impact of non-linearity.

There are also many interesting directions to extend our theoretical findings, including: proving tightness of the SDP relaxation for multiple input channels (we have formalized this as a plausible Conjecture 11); formally establishing the limiting behavior of gradient descent (without the caveats that we discussed); and exploring architectural features such as pooling or multiple layers.

We conclude our paper by positioning our results in relation to related work below:

Related Work Induced regularizers for different architectures has been studied directly and indirectly in the contexts of generalization and implicit regularization from gradient descent. Among recent work, Savarese et al. [SESS19] and Ongie et al. [OWSS20] directly studied the induced regularizer for infinite width two layer ReLU neural networks. As part of analysis of implicit bias from gradient descent, Gunasekar et al. [GLSS18b] characterized the induced regularizer for fully connected networks and for linear convolutional network with *full dimensional kernels*. Yun et al. [YKM21] extended Gunasekar et al. [GLSS18b] and showed a general connection for linear networks between implicit ℓ_1 norm minimization in an orthonormal basis and the existence of data-independent diagonalizations of the linear operator in each layer. [ZBHMS20] empirically demonstrated such differences.

In a work closely related to ours, Pilanci and Ergen [PE20], Ergen and Pilanci [EP20a; EP20b], and Sahiner et al. [SEPP20] study the induced regularizer of neural networks by looking at the bi-dual convex relaxation of the ℓ_2 regularized least squares loss. Phrased in the context of our work, their results on linear convolutional networks show that for networks with a single input channel, *if the number of output channels (or width) C is sufficiently large*, then the resulting induced regularizer is convex function that is independent of C, *i.e.,* the independence of C holds *after* C is above a certain large finite value. In contrast, we show the induced regularizer independent of the number of output channels *regardless of the number of output channels* and further that it is given by an SDP. Moreover these prior works focus on networks with a single input channel, while we extend our results to the networks with multi-channel inputs.

Finally, we mention the rich literature on work connecting ℓ_2 norm control of weights with explicit regularization [KH91; WLLM19] and implicit regularization from the gradient descent [NTS15; ZBHRV17; BFT17; GWBNS17; SHNGS18; GLSS18a; GLSS18b; JT18; JT19; NGLSS19; LL20; JT20; LL20]. While the implicit regularization from gradient descent trajectory is not always connected to ℓ_2 norm for regression (for example, see counter examples in [DFKL20; RC20]), the connection is prominent in many settings of interest including the asymptotic solution of minimizing logistic loss in homogeneous models [NGLSS19; LL20].

References

[Bar96] Peter L. Bartlett. “For Valid Generalization the Size of the Weights is More Important than the Size of the Network”. In: *Advances in Neural Information Processing Systems*. 1996.

[BFT17] Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. “Spectrally-normalized margin bounds for neural networks”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 6241–6250.

[BM02] Peter L. Bartlett and Shahar Mendelson. “Rademacher and Gaussian complexities: Risk bounds and structural results”. In: *Journal of Machine Learning Research* (2002), pp. 463–482.

[DFKL20] Assaf Dauber, Meir Feder, Tomer Koren, and Roi Livni. “Can Implicit Bias Explain Generalization? Stochastic Convex Optimization as a Case Study”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020.

[EP20a] Tolga Ergen and Mert Pilanci. “Convex Duality of Deep Neural Networks”. In: *CoRR* abs/2002.09773 (2020). arXiv: `2002.09773`. URL: `https://arxiv.org/abs/2002.09773`.

[EP20b] Tolga Ergen and Mert Pilanci. “Implicit Convex Regularizers of CNN Architectures: Convex Optimization of Two- and Three-Layer Networks in Polynomial Time”. In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. 2020.

[GLSS18a] Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. “Characterizing Implicit Bias in Terms of Optimization Geometry”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2018, pp. 1827–1836.

[GLSS18b] Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nati Srebro. “Implicit Bias of Gradient Descent on Linear Convolutional Networks”. In: *Advances in Neural Information Processing Systems*. 2018.

[GWBNS17] Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. “Implicit Regularization in Matrix Factorization”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 6151–6159.

[JT18] Ziwei Ji and Matus Telgarsky. “Risk and parameter convergence of logistic regression”. In: *CoRR* abs/1803.07300 (2018).

[JT19] Ziwei Ji and Matus Telgarsky. “Gradient descent aligns the layers of deep linear networks”. In: *International Conference on Learning Representations (ICLR)*. 2019.

[JT20] Ziwei Ji and Matus Telgarsky. “Directional convergence and alignment in deep learning”. In: *Advances in Neural Information Processing Systems*. 2020.

[KH91] Anders Krogh and John A. Hertz. “A Simple Weight Decay Can Improve Generalization”. In: *Advances in Neural Information Processing Systems*. 1991, pp. 950–957.

[Kri09] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Tech. rep. 2009.

[LC10] Yann LeCun and Corinna Cortes. “MNIST handwritten digit database”. In: (2010). URL: `http://yann.lecun.com/exdb/mnist/`.

[LL20] Kaifeng Lyu and Jian Li. “Gradient Descent Maximizes the Margin of Homogeneous Neural Networks”. In: *International Conference on Learning Representations (ICLR)*. 2020.

[NGLSS19] Mor Shpigel Nacson, Suriya Gunasekar, Jason D. Lee, Nathan Srebro, and Daniel Soudry. “Lexicographic and Depth-Sensitive Margins in Homogeneous and Non-Homogeneous Deep Models”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2019, pp. 4683–4692.

[NTS15] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. “In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning”. In: *International Conference on Learning Representations (ICLR), Workshop Track Proceedings*. 2015.

[OWSS20] Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. “A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case”. In: *International Conference on Learning Representations (ICLR)*. 2020.

[PE20] Mert Pilanci and Tolga Ergen. “Neural Networks are Convex Regularizers: Exact Polynomial-time Convex Optimization Formulations for Two-layer Networks”. In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 7695–7705. URL: `http://proceedings.mlr.press/v119/pilanci20a.html`.

[RC20] Noam Razin and Nadav Cohen. “Implicit Regularization in Deep Learning May Not Be Explainable by Norms”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020.

[RS05] Jason D. M. Rennie and Nathan Srebro. “Fast maximum margin matrix factorization for collaborative prediction”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2005, pp. 713–719.

[SEPP20] Arda Sahiner, Tolga Ergen, John M. Pauly, and Mert Pilanci. “Vector-output ReLU Neural Network Problems are Copositive Programs: Convex Analysis of Two Layer Networks and Polynomial-time Algorithms”. In: *CoRR* abs/2012.13329 (2020). arXiv: `2012.13329`. URL: `https://arxiv.org/abs/2012.13329`.

[SESS19] Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. “How do infinite width bounded norm networks look in function space?” In: *Proceedings of the Conference on Learning Theory (COLT)*. 2019, pp. 2667–2690.

[SHNGS18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. “The Implicit Bias of Gradient Descent on Separable Data”. In: *Journal of Machine Learning Research* (2018).

[WLLM19] Colin Wei, Jason D. Lee, Qiang Liu, and Tengyu Ma. “Regularization Matters: Generalization and Optimization of Neural Nets v.s. their Induced Kernel”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 9709–9721.

[YKM21] Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. “A Unifying View on Implicit Bias in Training Linear Neural Networks”. In: *International Conference on Learning Representations (ICLR)*. 2021.

[ZBHMS20] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Michael C. Mozer, and Yoram Singer. "Identity Crisis: Memorization and Generalization Under Extreme Overparameterization". In: *International Conference on Learning Representations (ICLR)*. 2020.

[ZBHRV17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations (ICLR)*. 2017.

A Experiments for gradient descent

We run our experiments on two layer linear convolutional networks on a subset of MNIST dataset [LC10] as well a subset of the CIFAR-10 dataset [Kri09]. The input images in MNIST are of size 28×28 and have a single input channel. The input images in MNIST are of size 32×32 and have a 3 input channels. We apply 2D convolutions with kernel sizes $K = (K_1, K_2)$ and circular padding for image inputs. We consider binary classification task for both datasets. For MNIST, we predict digits 0 and 1 in MNIST using a balanced sub-sampling of 128 samples as training data, which ensures linear separability. For CIFAR-10, we predict classes “automobile” and “dog” using a balanced sub-sampling of 256 samples as training data, which ensures linear separability. The initialization scale was taken to be 0.001.

We train our network using gradient descent on exponential loss and run gradient descent until the training loss is 10^{-6}. The initialization scale is taken to be 0.001 in order to reduce the variance arising from the randomness of initialization. In order to compare the predictors across different architectures, we normalize the weights learned by gradient descent $\mathbf{U}, \mathbf{V}$ such that the linear predictor $w(\mathbf{U}, \mathbf{V})$ realized by the trained networks has unit margin on the training dataset (*i.e.,* $y\langle w(\mathbf{U}, \mathbf{V}), \mathbf{x}\rangle \geq 1$ for all training samples $(\mathbf{x}, y)$). Note that for homogeneous models, such positive scaling of weights *does not* change the classification boundary of the learned model.[5]

A.1 Impact of the number of channels on MNIST

For networks with a single-input channel, Hypothesis 1 would imply that $\mathcal{R}_{K,C,R}(w(\mathbf{U}, \mathbf{V}))$ is invariant in the number of output channels C regardless of C. To demonstrate this, we repeat the experimental setup on 28×28 MNIST images on networks with multiple output channels $C \in \{1, 2, 4, 8\}$ and across different kernel sizes. As described earlier, we scale the weights learned by gradient descent $\mathbf{U}, \mathbf{V}$ such that the linear predictors $w(\mathbf{U}, \mathbf{V}) = \mathbf{w}_{\text{GD}}$ have unit margin on training data. Since it is difficult to directly compute $\mathcal{R}_{K,C,R}(\mathbf{w}_{\text{GD}})$, we turn to an approximation. In particular, we compute $\widehat{\mathcal{R}}_{K,C}(w(\mathbf{U}, \mathbf{V})) := \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2$. Strictly speaking, this is only an *upper bound* on the induced regularizer $\mathcal{R}_{K,C,R}(w(\mathbf{U}, \mathbf{V}))$.[6]

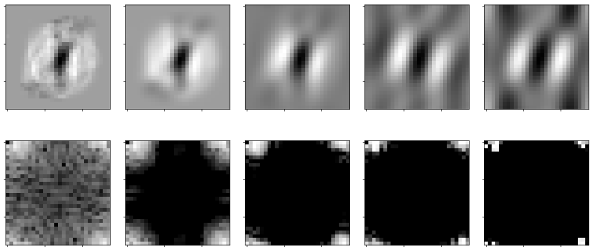
Table 1 shows $\widehat{\mathcal{R}}_{K,C}(w(\mathbf{U}, \mathbf{V}))$ across different values of K and C. We see that for each kernel size, the differences in $\widehat{\mathcal{R}}_{K,C}(w(\mathbf{U}, \mathbf{V}))$ across different settings of C are minimal and are usually smaller than the standard deviation for fixed settings of C. This suggests that the induced regularizer is indeed invariant to the number of output channels, thus providing evidence for Hypothesis 1 in the case of a single input channel.

C	$K = (1,1)$	$K = (3,3)$	$K = (8,8)$	$K = (16,16)$	$K = (28,28)$
1	$10.28 \pm 2.34 \times 10^{-5}$	$4.50 \pm 1.51 \times 10^{-3}$	$3.32 \pm 5.35 \times 10^{-2}$	$3.15 \pm 5.81 \times 10^{-2}$	$2.84 \pm 1.16 \times 10^{-1}$
2	$10.28 \pm 2.00 \times 10^{-5}$	$4.50 \pm 1.06 \times 10^{-3}$	$3.30 \pm 3.13 \times 10^{-2}$	$3.10 \pm 2.63 \times 10^{-2}$	$2.79 \pm 1.27 \times 10^{-1}$
4	$10.28 \pm 1.00 \times 10^{-5}$	$4.50 \pm 7.48 \times 10^{-4}$	$3.30 \pm 2.25 \times 10^{-2}$	$3.10 \pm 3.33 \times 10^{-2}$	$2.77 \pm 8.20 \times 10^{-2}$
8	$10.28 \pm 7.83 \times 10^{-5}$	$4.50 \pm 6.25 \times 10^{-4}$	$3.29 \pm 1.68 \times 10^{-2}$	$3.11 \pm 3.27 \times 10^{-2}$	$2.72 \pm 7.31 \times 10^{-2}$

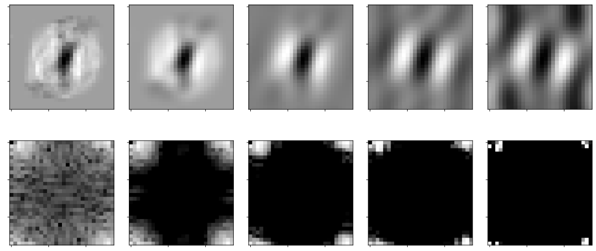
Table 1: $\widehat{\mathcal{R}}_{K,C}(w(\mathbf{U}, \mathbf{V})) = \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2$ of the predictor learned by gradient descent on linear convolutional networks with different number of output channels C and kernel sizes K on the MNIST task. We show the mean over 10 trials as well as the standard deviations are also shown.

[5] The code is available at `https://github.com/mjagadeesan/inductive-bias-multi-channel-CNN`.

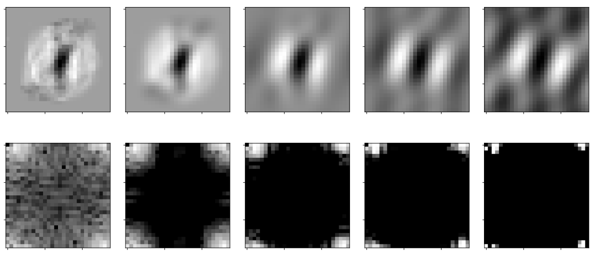
[6] For $K = 1$ and $K = D$, we verified that the estimate is close to tight by computing the ℓ_2 and ℓ_1 norms of Fourier transform of the predictor, respectively.



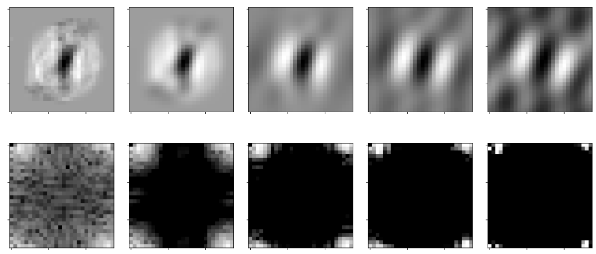
(a) $C = 1$



(b) $C = 2$



(c) $C = 4$



(d) $C = 8$

Figure 1: Linear predictors learned by two layer linear convolutional network for the task of classifying digits 0 and 1 in MNIST. The sub-figures depict predictors learned by using gradient descent on the exponential loss for overparameterized networks with $C = 1, 2, 4$ and kernel sizes $K \in \{(1,1),(3,3),(8,8),(16,16),(28,28)\}$ (left to right). The top row in each sub–figure is the signal domain representation $w(\mathbf{U},\mathbf{V})$, and the bottom row is the Fourier domain representation $\widehat{w}(\mathbf{U},\mathbf{V})$.

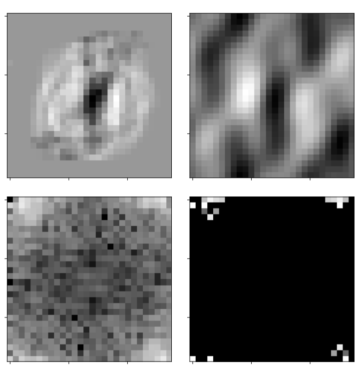


Figure 2: Explicit $\mathcal{R}_{K,C}$ margin predictor on sampled MNIST dataset for kernel sizes $K \in \{(1,1),(28,28)\}$ (left to right). The top row in each sub–figure is the signal domain representation $w(\mathbf{U},\mathbf{V})$, and the bottom row is the Fourier domain representation $\widehat{w}(\mathbf{U},\mathbf{V})$.

Invariance of learned predictors to C. While Hypothesis 1 primarily pertains to the behavior of the induced regularizer, it also suggests that the predictor $\mathbf{w}_{\mathrm{GD}}$ will also be independent of the number of channels so long as K is strictly less than D. If we overlook the caveats, we expect the gradient descent to implicitly learn a max–$\mathcal{R}_{K,C}$ margin predictor: $\min_{\mathbf{w}} \mathcal{R}_{K,C}(\mathbf{w})$ s.t., $\forall_n y_n\langle \mathbf{w}, \mathbf{x}_n\rangle \geq 1$. For $K < D$, our theoretical findings suggest that the induced regularizer is a norm interpolating between the ℓ_2 and ℓ_1 norms. This would mean that there is a *unique* global minimizer, and thus we would expect that $\mathbf{w}_{\mathrm{GD}}$ to be invariant to C.

To empirically validate this, we show the learned linear predictors for $C = \{1, 2, 4\}$ in Figure 1. We observe that the linear predictors indeed visually appears to be invariant across different settings of C for all for kernel sizes $K < D$. For $K = (28, 28)$, there appear to be differences in the predictors—this likely arises from the fact that there are multiple linear predictors that minimize the ℓ_1 norm on the dataset. We nonetheless emphasize that the *induced regularizer* still appears to be invariant in this case, although the predictors are not.

A.1.1 Non-linear networks with ReLU activation

Although our theoretical results are restricted to networks with linear activations, it is nevertheless interesting to evaluate if our conclusions lead to useful heuristics for networks with non-linearity. As a simple demonstration, we repeat our experiment on MNIST on two-layer convolutional networks with ReLU non-linearity with and without bias parameters (*i.e.,* networks with a convolution layer, followed by ReLU layer, followed by linear layer).[7] As before, we first scale the weights learned by gradient descent such that the resulting predictor f_{GD} has unit margin on training data. We then consider the representation cost $\mathcal{R}_{\Phi_{K,C}}(f_{\mathrm{GD}})$, as per equation (2), given by the minimum ℓ_2 norm of the weights needed to realize f. We consider the approximation of $\mathcal{R}_{\Phi_{K,C}}(f_{\mathrm{GD}})$ given by $\widehat{\mathcal{R}}_{\Phi_{K,C}}(f_{\mathrm{GD}}) := \|\mathbf{U}\|_2^2 + \|\mathbf{V}\|_2^2$ where $\mathbf{U}$ and $\mathbf{V}$ are the weights learned by gradient descent. (As before, strictly speaking, this is only an *upper bound* on the representation cost $\mathcal{R}_{\Phi_{K,C}}(f_{\mathrm{GD}})$.)

Table 2 and Table 3 show $\widehat{\mathcal{R}}_{\Phi_{K,C}}(f_{\mathrm{GD}}) := \|\mathbf{U}\|_2^2 + \|\mathbf{V}\|_2^2$ across different settings of C and K, for networks with no bias as well as networks with bias parameters on both the convolution layer and the fully connected layer.[8] Like in the case of linear convolutional neural networks, $\widehat{\mathcal{R}}_{\Phi_{K,C}}(f_{\mathrm{GD}})$ is consistent across different settings of C. This suggests that the implicit bias from gradient might result in predictors that are independent of the number of output channels, even when there is a ReLU layer, and Hypothesis 1 might hold in much more generality than the scope of theoretical findings.

C	$K : (1,1)$	$K : (3,3)$	$K : (8,8)$	$K : (16,16)$	$K : (28,28)$
1	11.412	5.160	3.998	3.785	3.520
2	11.413	5.155	3.964	3.721	3.539
4	11.414	5.153	3.966	3.719	3.448
8	11.415	5.156	3.971	3.738	3.498

Table 2: $\widehat{\mathcal{R}}_{K,C}(f_{\mathrm{GD}}) = \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2$ of the predictor learned by gradient descent on ReLU convolutional networks without bias parameters, with different number of output channels C and kernel sizes K, on the MNIST task. The values shown are the medians taken over 5 trials.

We note that we observed that gradient descent sometimes leads to outliers where $\mathcal{R}_{\Phi_{K,C}}(f_{\mathrm{GD}})$ is very large. For example, when $K = 1$, the values of $\mathcal{R}_{\Phi_{K,C}}(f)$ are $[11.412, 11.412, 109.471, 11.412, 11.412]$,

[7] The initialization scale was taken to be 0.005 for networks without bias parameters and 0.01 for networks with bias parameters.

[8] We note that the representation cost includes the magnitude of the *weights* but not the magnitude of the *biases*.

C	$K:(1,1)$	$K:(3,3)$	$K:(8,8)$	$K:(16,16)$	$K:(28,28)$
1	10.581	4.948	3.875	3.714	3.519
2	10.571	4.945	3.910	3.698	3.413
4	10.578	4.945	3.912	3.712	3.399
8	10.576	4.946	3.881	3.697	3.437

Table 3: $\widehat{\mathcal{R}}_{K,C}(f_{\mathrm{GD}}) = \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2$ of the predictor learned by gradient descent on ReLU convolutional networks with bias on both layers, with different number of output channels C and kernel sizes K, on the MNIST task. The values shown are the medians taken over 5 trials.

where 109.471 appears to be an outlier. We anticipate that this outlier arises because gradient descent converges to a stationary point, rather than a local minima, of the max-ℓ_2 margin problem in parameter space (see the discussion in Section 1.1). Since our goal is to investigate the behavior of gradient descent when it does lead to global minima of the $max - \mathcal{R}_\Phi$ margin problem, we compute the median so that these data points do not affect our estimate.

A.2 Impact of the number of channels on CIFAR-10

We carry out a similar investigation of Hypothesis 1 on the CIFAR-10 dataset for networks with 3-channel inputs. As discussed in Section 5, we expect that the induced regularizer is *not* independent of the number of output channels for $C < R$, but begins to exhibit invariance once $C \geq R$. On the CIFAR-10 dataset, where there are 3 input channels, we would expect to see invariance once $C \geq 3$.

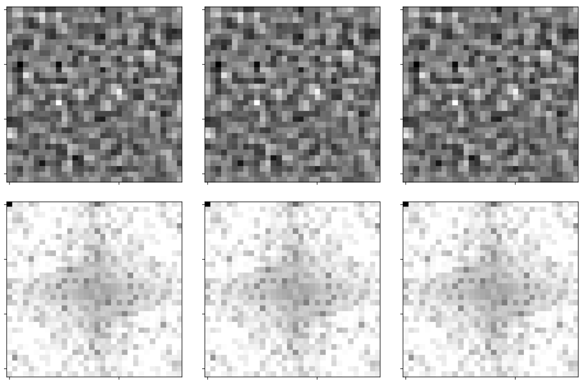
To demonstrate this, we repeat the experimental setup on 32×32 CIFAR-10 images on networks with multiple output channels $C \in \{1, 2, 4, 8\}$ and across different kernel sizes. As described earlier, we scale the weights learned by gradient descent $\mathbf{U}, \mathbf{V}$ such that the linear predictor $w(\mathbf{U}, \mathbf{V}) = \mathbf{w}_{\mathrm{GD}}$ has unit margin on training data. Since it is difficult to directly compute $\mathcal{R}_{K,C,R}(\mathbf{w}_{\mathrm{GD}})$, we turn to an approximation, as we did in the case of single-input channels. We compute $\widehat{\mathcal{R}}_{K,C,R}(w(\mathbf{U}, \mathbf{V})) := \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2$ which strictly speaking, this is only an *upper bound* on the induced regularizer $\mathcal{R}_{K,C,R}(w(\mathbf{U}, \mathbf{V}))$.

Table 4 shows $\widehat{\mathcal{R}}_{K,C,3}(w(\mathbf{U}, \mathbf{V}))$ across different values of K and C. We see that for each kernel size, the differences in $\widehat{\mathcal{R}}_{K,C,3}(w(\mathbf{U}, \mathbf{V}))$ across different settings of C are minimal, as long as $C \geq 3$ (and often, even when $C \geq 2$). This suggests that the induced regularizer is indeed invariant to the number of output channels when $C \geq 3$, thus providing evidence for Hypothesis 1 in the case of multiple input channels. Moreover, there are non-trivial differences in $\widehat{\mathcal{R}}_{K,C,3}(w(\mathbf{U}, \mathbf{V}))$ for $C = 1$ and larger C—this aligns with our theoretical findings in Section 5 that the induced regularizer does depend on C when it is below R.

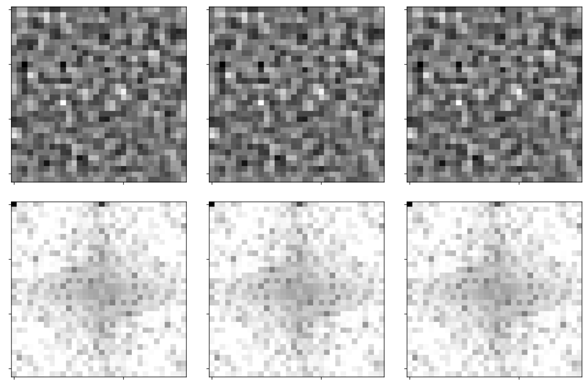
While Hypothesis 1 primarily pertains to the behavior of the induced regularizer, we would also expected that there is a unique global minimizer in most cases, for reasons similar to for the single input channel case. Thus, we would expect $\mathbf{w}_{\mathrm{GD}}$ to be invariant to C as long as $C \geq R = 3$. To empirically validate this, we show the learned linear predictors for $C = \{1, 2, 3, 4, 8\}$ in Figures 3-5. We observe that the linear predictors indeed visually appears to be invariant across different settings of C.

A.3 Varying kernel sizes

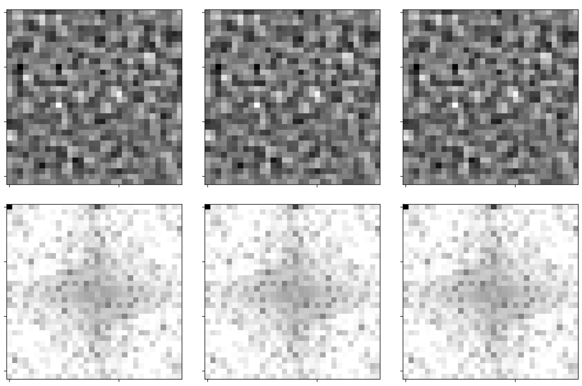
While the number of output channels has little influence on the induced regularizer of the learned predictors, we show that the kernel size can have significant impact, which aligns with our theoretical findings.



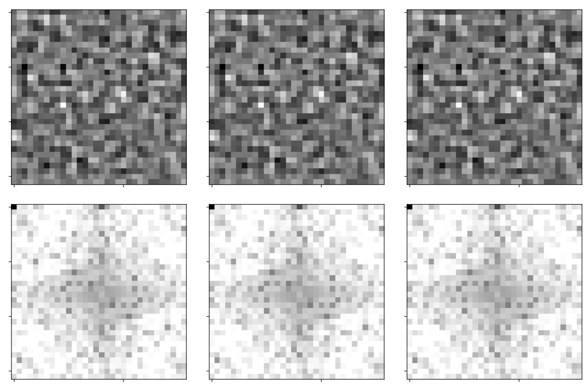
(a) $C = 1$



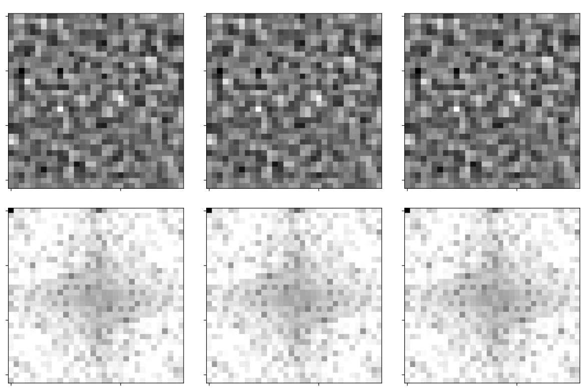
(b) $C = 2$



(c) $C = 3$

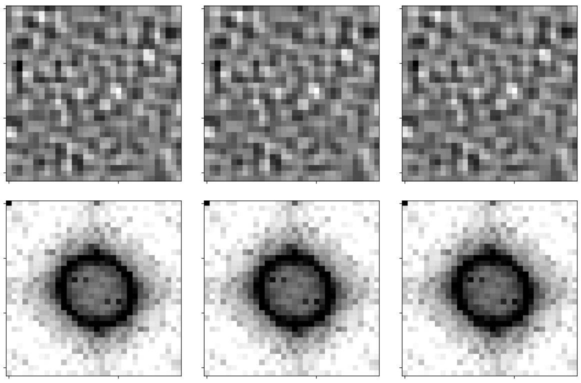


(d) $C = 4$

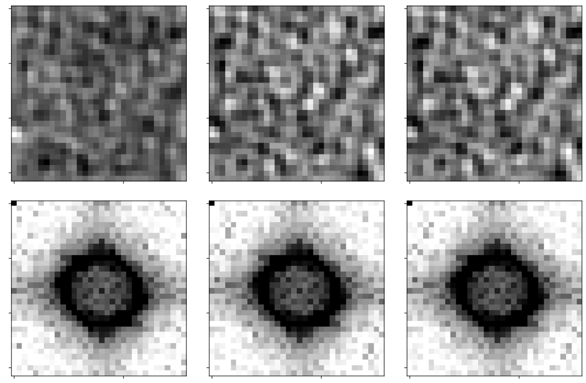


(e) $C = 8$

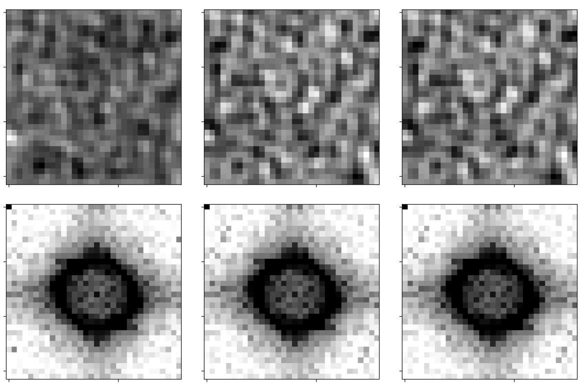
Figure 3: Linear predictors learned by two layer linear convolutional network on CIFAR-10 task. The sub-figures depict predictors learned by using gradient descent on the exponential loss for overparameterized networks with $C \in \{1, 2, 3, 4, 8\}$ and kernel size $K = (1, 1)$. The top row in each sub–figure is the signal domain representation $w(\mathbf{U}, \mathbf{V})$, and the bottom row is the Fourier domain representation $\widehat{w}(\mathbf{U}, \mathbf{V})$.



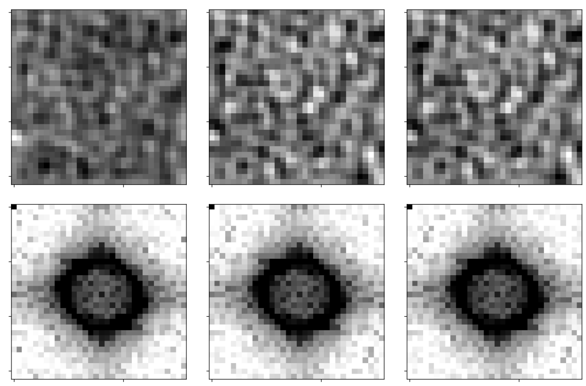
(a) $C = 1$



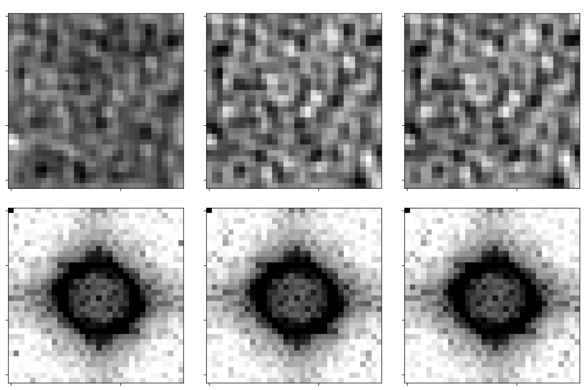
(b) $C = 2$



(c) $C = 3$

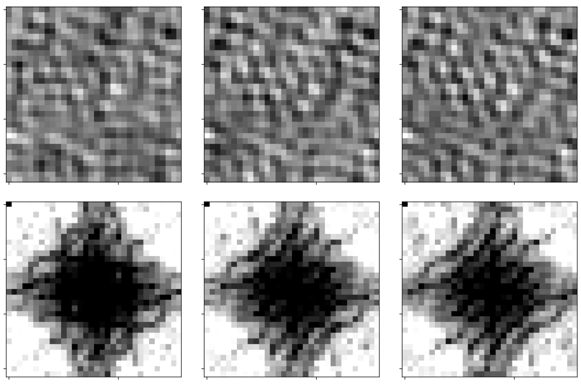


(d) $C = 4$

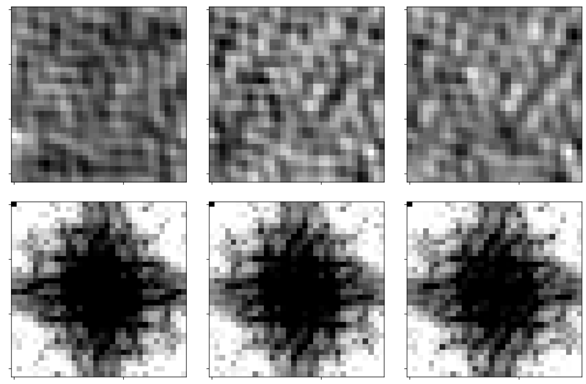


(e) $C = 8$

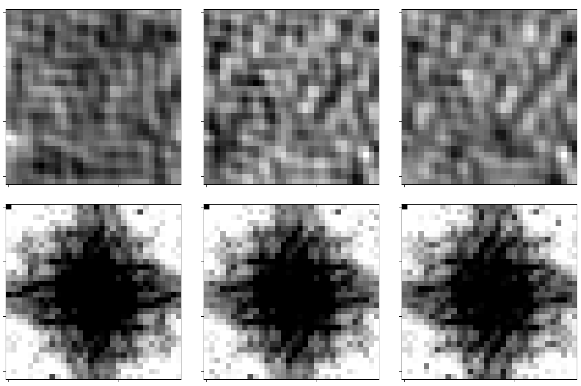
Figure 4: Linear predictors learned by two layer linear convolutional network on CIFAR-10 task. The sub-figures depict predictors learned by using gradient descent on the exponential loss for overparameterized networks with $C \in \{1, 2, 3, 4, 8\}$ and kernel size $K = (3, 3)$. The top row in each sub–figure is the signal domain representation $w(\mathbf{U}, \mathbf{V})$, and the bottom row is the Fourier domain representation $\widehat{w}(\mathbf{U}, \mathbf{V})$.



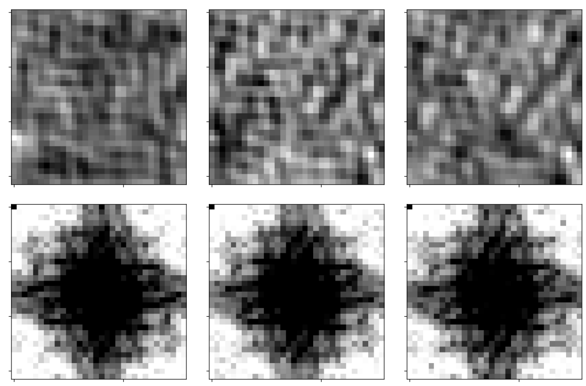
(a) $C = 1$



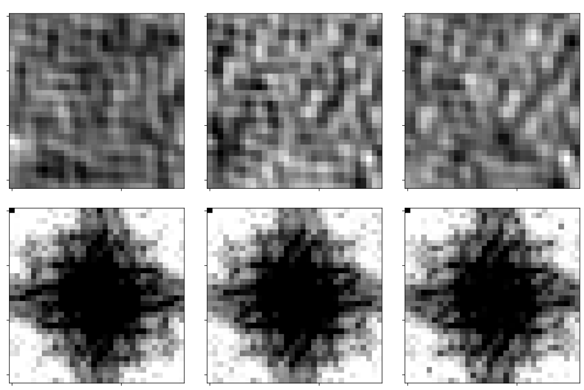
(b) $C = 2$



(c) $C = 3$



(d) $C = 4$



(e) $C = 8$

Figure 5: Linear predictors learned by two layer linear convolutional network on CIFAR-10 task. The sub-figures depict predictors learned by using gradient descent on the exponential loss for overparameterized networks with $C \in \{1, 2, 3, 4, 8\}$ and kernel size $K = (8, 8)$. The top row in each sub–figure is the signal domain representation $w(\mathbf{U}, \mathbf{V})$, and the bottom row is the Fourier domain representation $\widehat{w}(\mathbf{U}, \mathbf{V})$.

C	$K=(1,1)$	$K=(3,3)$	$K=(8,8)$	$K=(20,20)$
1	246.04	215.21	202.27	131.16
2	246.26	182.77	168.40	124.66
3	245.98	182.80	165.32	123.56
4	246.29	182.83	164.50	123.37
8	245.58	182.82	164.86	123.59

Table 4: $\widehat{\mathcal{R}}_{K,C,3}(w(\mathbf{U},\mathbf{V})) = \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2$ of the predictor learned by gradient descent on linear convolutional networks with different number of output channels C and kernel sizes K on the CIFAR-10 task.

A.3.1 Effect of kernel size on MNIST

Our theoretical findings in Section 4.1 suggest that the induced regularizer interpolates between ℓ_2 and ℓ_1 norms in the Fourier domain. The ℓ_2 regularization of $\mathcal{R}_{1,1}(\mathbf{w})$ does not induce sparse solutions, while the ℓ_1 regularization of $\mathcal{R}_{D,1}(\mathbf{w})$ promotes sparsity in the Fourier basis. This would suggest the following: *larger kernel sizes induce sparsity in the frequency domain.*

Explicit optimal solutions for $K=(1,1)$ and $K=(D,D)$. First, to illustrate in the extreme cases of $K=1$ and $K=D$, we explicitly compute the $\mathcal{R}_{K,C}$ margin predictor:

$$\min_{\mathbf{w}} \mathcal{R}_{K,C}(\mathbf{w}) \text{ s.t., } \forall_n y_n \langle \mathbf{w}, \mathbf{x}_n \rangle \geq 1$$

on the dataset using the closed-form solutions for the induced regularizer in these special cases. In Figure 2, we show resulting optimal solutions for $K=D$ (a minimum ℓ_1 solution) and $K=1$ (a minimum ℓ_2 solution). While the solution for $K=(1,1)$ exhibits no sparsity in the frequency domain, the solution for $K=(D,D)$ exhibits significant sparsity.

The corresponding values of $\mathcal{R}_{K,C}$ are 9.32 for $K=D$ and 2.10 for $K=1$. We note that the induced regularizer do not exactly match those computed on the $w(\mathbf{U},\mathbf{V})$ from gradient descent—this is because the convergence of these values can be quite slow. Moreover, for $K=D$, the difference in the predictors likely stems from the minimum ℓ_1-norm solution being non-unique. We nonetheless show that the qualitative findings apply to gradient descent, despite the fact that the limiting values have not been reached.

Extension to gradient descent. Consider networks with one output channel $C=1$ and compute $w(\mathbf{U},\mathbf{V})$ learned by gradient descent for networks with different kernel sizes in Figure 1-(a). Notice in the frequency domain plots that the predictor learned with kernel size $K=(1,1)$ is not sparse, the predictor learned with $K=(3,3)$ already starts to exhibit some sparsity, and the linear predictor learned with $K=(28,28)$ is highly sparse in the frequency domain.

Since sparsity in the frequency domain promotes a patterned structure in the signal domain, we explore the qualitative behavior of large kernel sizes in the signal domain in more depth. To do this, we construct an augmented version of the dataset with 112×112 dimensional images where the top-left 28×28 region is the original image, while the remaining space is all 0s. Figure 6 shows the linear predictors learned by running gradient descent on single output channel networks with different kernel sizes. As K increases, the nonzero region of the predictor becomes larger, eventually encompassing the full 112×112 dimensional space. For large kernel sizes, we can visually see that the predictors are composed of repetitions of a pattern. This is suggestive of a restricted form of periodic translation invariance, where the shift size aligns with the size of the patterns.

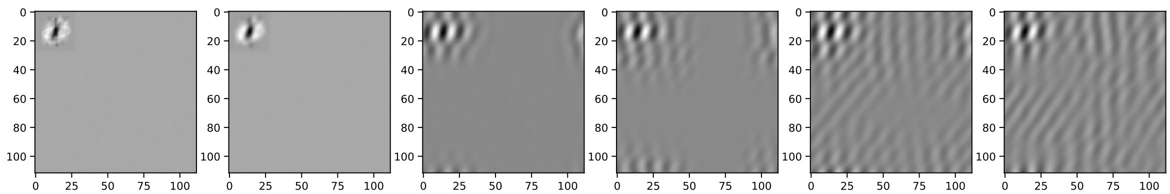


Figure 6: Linear predictors learned by gradient descent on single output channel networks over an augmented input space with kernel sizes $K \in \{(1,1),(3,3),(27,27),(45,45),(65,65),(84,84)\}$ (left to right). The input images from the MNIST dataset are augmented by padding with zeros to obtain an image of size 112×112 with the signal present only in the top-left 28×28 block.

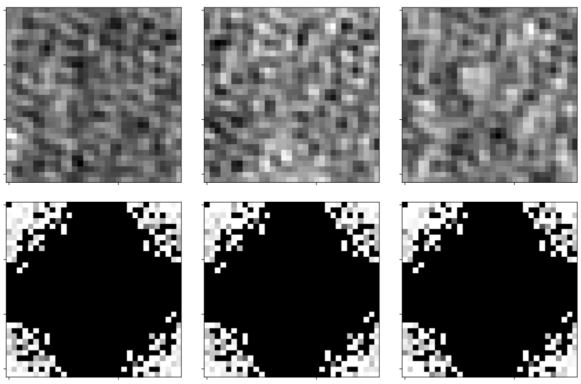
A.3.2 Effect of kernel size on CIFAR-10

We now examine the role of kernel size for multi-channel networks on CIFAR-10. Our theoretical findings in Section 5 suggest that the induced regularizer interpolates between the nuclear norm for $K = 1$ and the $\ell_{2,1}$ norm for $K = D$. This would again suggest the following: *larger kernel sizes induce sparsity in the frequency domain.* The behavior across input channels, however, is more nuanced. Both of these norms favor similarities across different input channels (with the effect intuitively stronger for $K = 1$ since the nuclear norm is closely related to rank). We explore both of these effects in the following experiments.

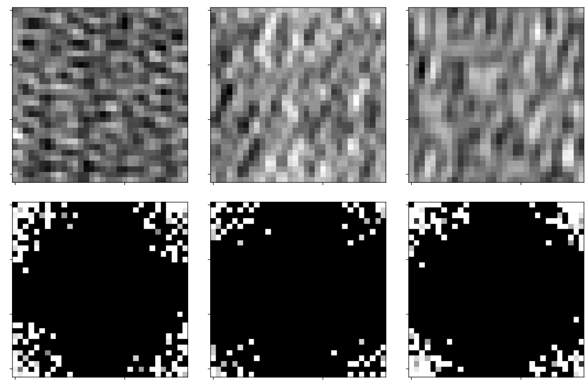
Explicit optimal solutions for $K = (D, D)$. First, to illustrate in the extreme case of $K = D$, we explicitly compute the $\mathcal{R}_{K,C}$ margin predictor: $\min_{\mathbf{w}} \mathcal{R}_{K,C}(\mathbf{w})$ s.t., $\forall_n y_n \langle \mathbf{w}, \mathbf{x}_n \rangle \geq 1$ on the dataset using the closed-form solutions for the induced regularizer in these special cases. In Figure 7, we show resulting optimal solutions for $K = D$ (a minimum $\ell_{2,1}$ solution) along with the optimal $\ell_{1,1}$ solution. We visually see that the $\ell_{2,1}$ solution favors similarity across input channels at the expense of greater sparsity in the frequency domain.

The corresponding values of $\mathcal{R}_{D,C}$ for the minimum $\ell_{2,1}$ norm solution is 82.85 for $K = (D, D)$.) As for the single-input channel case, we note that the induced regularizer do not exactly match the value of 114.85 of the induced regularizer computed on the $w(\mathbf{U}, \mathbf{V})$ from gradient descent—this is because the convergence of the induced regularizer is known to be slow. We nonetheless show that the qualitative findings apply to gradient descent, despite the fact that the limiting value of induced regularizer has not been reached.

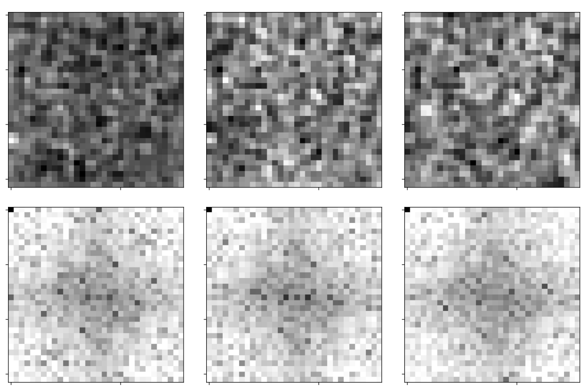
Extension to gradient descent. Consider networks with one output channel $C = 3$ and compute $w(\mathbf{U}, \mathbf{V})$ learned by gradient descent for networks with different kernel sizes in Figure 8. First, the higher kernel does indeed favor sparsity in the frequency domain, as in the single-input channel case. Nontrivial sparse structures can be observed even for $K = 3$. Next, we discuss the predictor across different input channels. Let's first focus on the extreme case of $K = 1$. For sake of comparison, we show the explicit minimum ℓ_2 (Frobenius) predictor in Figure 7 (note that this is *not* the $\mathcal{R}_{1,C}$ margin predictor because the induced regularizer is related to the nuclear norm, not the ℓ_2 norm). As expected, we see that the learned predictor has a greater degree of similarity across channels than the ℓ_2 predictor. For other kernel sizes, Figure 8 also shows some degree of similarity across input channels, although the differences appear to grow as K becomes larger.



(a) $\ell_{2,1}$

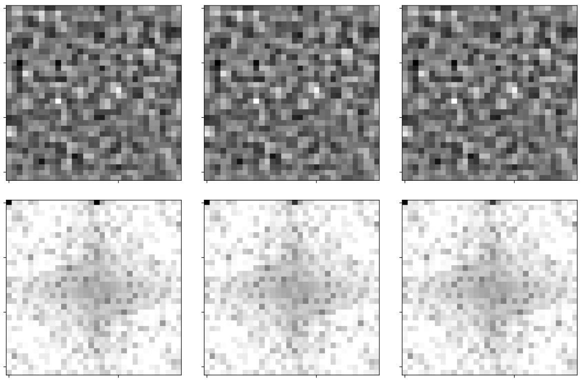


(b) ℓ_1

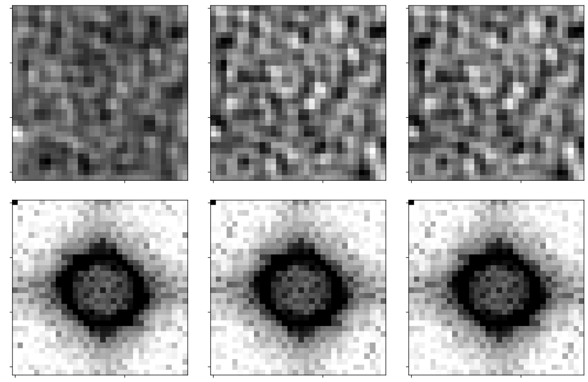


(c) ℓ_2

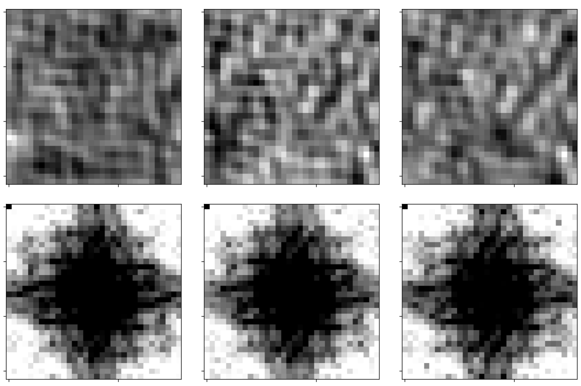
Figure 7: Explicit $\ell_{2,1}$, ℓ_1 and ℓ_2 margin predictors on sampled CIFAR-10 dataset. The top row in each sub–figure is the signal domain representation $w(\mathbf{U}, \mathbf{V})$, and the bottom row is the Fourier domain representation $\widehat{w}(\mathbf{U}, \mathbf{V})$



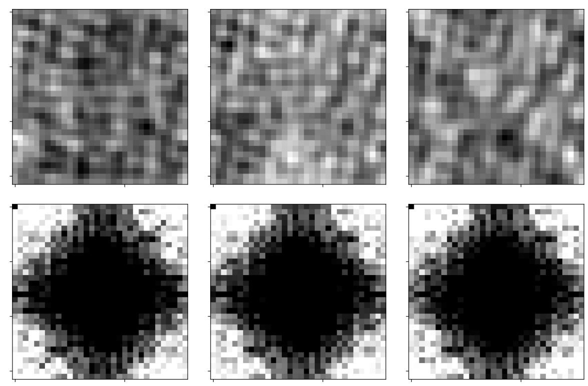
(a) $K = (1, 1)$



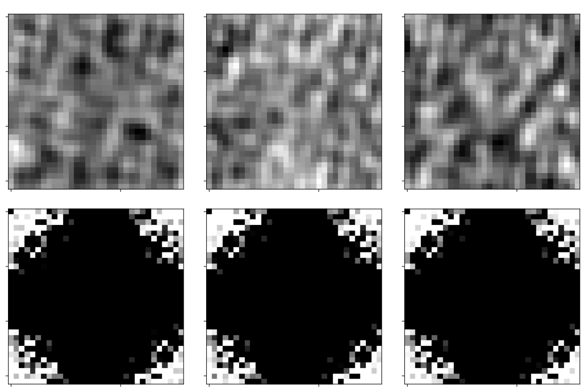
(b) $K = (3, 3)$



(c) $K = (8, 8)$



(d) $K = (20, 20)$



(e) $K = (32, 32)$

Figure 8: Linear predictors learned by two layer linear convolutional network on CIFAR-10 task. The sub-figures depict predictors learned by using gradient descent on the exponential loss for overparameterized networks with $C = 3$ and kernel sizes $K \in \{(1, 1), (3, 3), (8, 8), (20, 20), (32, 32)\}$. The top row in each sub–figure is the signal domain representation $w(\mathbf{U}, \mathbf{V})$, and the bottom row is the Fourier domain representation $\widehat{w}(\mathbf{U}, \mathbf{V})$.

B Proofs for Section 3: Induced regularizer in special cases

For a single channel convolutional network (*i.e.,* $C = 1$), we denote the weights in the first and second layer as $\mathbf{U} \in \mathbb{R}^K$ and $\mathbf{V} \in \mathbb{R}^D$, respectively. Recall that the D dimensional discrete Fourier transform of the weights $\mathbf{U}, \mathbf{V}$ and the linear predictor $\mathbf{w} \in \mathbb{R}^D$ are denoted as $\widehat{\mathbf{U}} = \mathbf{F}_K\mathbf{U}, \widehat{\mathbf{V}} = \mathbf{F}\mathbf{V}, \widehat{\mathbf{w}} = \mathbf{F}\mathbf{w}$, respectively. Moreover, the Fourier transform is normalized to be unitary such that for any $\mathbf{z} \in \mathbb{R}^K$, $\|\mathbf{z}\| = \|\widehat{\mathbf{z}}\|$ and $\mathbf{F}\overline{\mathbf{F}}^\top = \overline{\mathbf{F}}\mathbf{F}^\top = \mathbf{I}$. Thus, for all the quantities of interest, we use the ℓ_2 norm in signal domain interchangeably with ℓ_2 norm in the Fourier domain, *e.g.,* $\|\mathbf{U}\| = \|\widehat{\mathbf{U}}\|$, $\|\mathbf{V}\| = \|\widehat{\mathbf{V}}\|$, and $\|\mathbf{w}\| = \|\widehat{\mathbf{w}}\|$.

In the following proofs, we use the formulation of $\mathcal{R}_{K,C}(\mathbf{w})$ in eq. (7) for the case of $C = 1$ as:

$$\mathcal{R}_{K,1}(\mathbf{w}) = \min_{\mathbf{U}\in\mathbb{R}^K,\mathbf{V}\in\mathbb{R}^D} \|\widehat{\mathbf{U}}\|^2 + \|\widehat{\mathbf{V}}\|^2 \quad \text{s.t.,} \quad \widehat{\mathbf{U}} \odot \widehat{\mathbf{V}} = \widehat{\mathbf{w}}.$$

B.1 Proof of Lemma 2

Lemma 2 ($K = 1$)**.** *For any $\mathbf{w} \in \mathbb{R}^D$, it holds that $\mathcal{R}_{1,1}(\mathbf{w}) = 2\sqrt{D}\|\widehat{\mathbf{w}}\|_2 = 2\sqrt{D}\|\mathbf{w}\|_2$.*

Proof. This statement is trivially true for $\mathbf{w} = 0$, so it suffices to show this for $\mathbf{w} \neq 0$. When the $K = 1$, the first layer weight $\mathbf{U} \in \mathbb{R}^{1\times 1}$ is a scalar. Let this scalar be $\mathbf{U} = u \neq 0$. We then have $\widehat{\mathbf{U}} = \frac{1}{\sqrt{D}}[u, u, \ldots, u]$. Since $\widehat{\mathbf{U}} \odot \widehat{\mathbf{V}} = \widehat{\mathbf{w}}$, we have $\widehat{\mathbf{V}} = \frac{\sqrt{D}}{u}\widehat{\mathbf{w}}$. This means that $\mathcal{R}_{1,1}(\mathbf{w}) = \min_u \|\widehat{\mathbf{U}}\|^2 + \|\widehat{\mathbf{V}}\|^2 = \min_u u^2 + \frac{D}{u^2}\|\widehat{\mathbf{w}}\|_2^2$. By using the AM-GM inequality ($a^2 + b^2 \geq 2ab$), this is at most $2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$. Moreover, we can pick $u^2 = \sqrt{D}\|\widehat{\mathbf{w}}\|_2$ to achieve equality. □

B.2 Proof of Lemma 3

Lemma 3. *For any $\mathbf{w} \in \mathbb{R}^D$, it holds that:*

$$\mathcal{R}_{2,1}(\mathbf{w}) = 2\sqrt{D}\sqrt{\inf_{\alpha\in(-1,1)} \sum_{d=0}^{D-1} \frac{|\widehat{\mathbf{w}}[d]|^2}{1 + \alpha\cos(2\pi d/D)}}.$$

Proof of Lemma 3. We first note that for any $\mathbf{U} \in \mathbb{R}^K, \mathbf{V} \in \mathbb{R}^D$ we can re-scale the norms so that $\|\mathbf{U}\| = \|\mathbf{V}\|$ while satisfying the constraints of $\widehat{\mathbf{U}} \odot \widehat{\mathbf{V}} = \widehat{\mathbf{w}}$ in the definition of $\mathcal{R}_{K,1}(\mathbf{w})$. Further, such a scaling would be optimal for minimizing the ℓ_2 norm of weights based on AM-GM inequality that $\|\widehat{\mathbf{U}}\|^2 + \|\widehat{\mathbf{V}}\|^2 \geq 2\|\widehat{\mathbf{U}}\| \cdot \|\widehat{\mathbf{V}}\|$. Thus, in the rest of the proof, we consider the following equivalent formulation of $\mathcal{R}_{K,1}(\mathbf{w})$ as:

$$\mathcal{R}_{K,1}(\mathbf{w}) = \min_{\mathbf{U}\in\mathbb{R}^K,\mathbf{V}\in\mathbb{R}^D} 2\|\mathbf{U}\| \cdot \|\widehat{\mathbf{V}}\| \quad \text{s.t.,}\ \widehat{\mathbf{U}} \odot \widehat{\mathbf{V}} = \widehat{\mathbf{w}} \tag{15}$$

We see that for any $\mathbf{U}, \mathbf{V}$ satisfying the constraint in the above equation, we have: $\forall_{d\in\text{supp}(\widehat{\mathbf{w}})}$, it holds that $\widehat{\mathbf{V}}[d] = \frac{\widehat{\mathbf{w}}[d]}{\widehat{\mathbf{U}}[d]}$ (where $\text{supp}(\widehat{\mathbf{w}}) = \{d \in [D] : |\widehat{\mathbf{w}}| \neq 0\}$). Moreover, at an optimal solution, it is easy to see that $\widehat{\mathbf{V}}[d] = 0 \iff \widehat{\mathbf{w}}[d] = 0$.

Let $\mathbf{U} = [c_0, c_1]$. This means that the objective can be written as

$$2\|\mathbf{U}\|\|\widehat{\mathbf{V}}\| = 2\|\mathbf{U}\|\sqrt{\sum_{d=0}^{D-1} |\widehat{\mathbf{V}}[d]|^2} = 2\|\mathbf{U}\|\sqrt{\sum_{d\in\text{supp}(\widehat{\mathbf{w}})} |\widehat{\mathbf{V}}[d]|^2} = 2\sqrt{c_0^2 + c_1^2}\sqrt{\sum_{d\in\text{supp}(\widehat{\mathbf{w}})} \frac{|\widehat{\mathbf{w}}[d]|^2}{|\widehat{\mathbf{U}}[d]|^2}}. \tag{16}$$

We write the second term in terms of the signal domain representation of c_0 and c_1. We see that the Fourier transform of $\mathbf{U}$ is given by $\widehat{\mathbf{U}}[d] = \frac{1}{\sqrt{D}}\left(c_0 + c_1 \mathrm{e}^{-2\pi i d/D}\right) = (c_0 + c_1\cos(-2\pi i d/D)) + i\,(c_1 \sin(-2\pi i d/D))$. We thus have that:

$$\begin{aligned}
|\widehat{\mathbf{U}}[d]|^2 &= \frac{1}{D}\left[(c_0 + c_1\cos(-2\pi i d/D))^2 + (c_1\sin(-2\pi i d/D))^2\right] \\
&= \frac{1}{D}\left(c_0^2 + c_1^2 + 2c_0c_1\cos(-2\pi i k/D)\right). \\
&= \frac{1}{D}(c_0^2 + c_1^2)\left(1 + \frac{2c_0c_1}{c_0^2+c_1^2}\cos(-2\pi i d/D)\right).
\end{aligned} \tag{17}$$

Let $\alpha = \frac{2c_0c_1}{c_0^2+c_1^2}$. Plugging eq. (17) back into the objective in eq. (16) and using $\cos(z) = \cos(-z)$, we get that for any $\mathbf{U}, \mathbf{V}$ satisfying the constraints in the computation of $\mathcal{R}_{2,1}(\mathbf{w})$, the objective is in the desired formulation:

$$2\|\mathbf{U}\|\|\widehat{\mathbf{V}}\| \;= 2\sqrt{D}\sqrt{\sum_{d\in\mathrm{supp}(\widehat{\mathbf{w}})} \frac{|\widehat{\mathbf{w}}[d]|^2}{1+\alpha\cos(2\pi i d/D)}}. \tag{18}$$

Let us now consider the domain of α, which is the only unknown in the above equation. Observe that for any c_0, c_1, $\frac{2c_0c_1}{c_0^2+c_1^2} \in [-1,1]$. Moreover, any $\alpha \in [-1,1]$ be realized by some values of c_0 and c_1. Thus all $\alpha \in [-1,1]$ are valid. Here we further remark that the denominator in eq. (18) is zero if and only if $\widehat{\mathbf{U}}[d] = 0$ for any $d \in D$. However, this can only happen if $\widehat{\mathbf{w}}[d] = 0$ as otherwise the constraints $\widehat{\mathbf{U}} \odot \widehat{\mathbf{V}} = \mathbf{w}$ is not satisfied for any $\mathbf{V}$. We can thus, minimize the RHS of eq. (18) over $\alpha \in [-1,1]$ to obtain $\mathcal{R}_{2,1}(\mathbf{w})$.

If we include the terms corresponding to $d \notin \mathrm{supp}(\widehat{\mathbf{w}})$ in the summation in eq. (18), there is a technical condition than can lead to $0/0$ terms in end cases of $\alpha = 1$ (when $\widehat{\mathbf{w}}[0] = 0$) and $\alpha = -1$ (when $\widehat{\mathbf{w}}[D/2] = 0$). To avoid this technicality, we consider the infimum over $\alpha \in (-1,1)$ rather than minimum over $\alpha \in [-1,1]$. This is equivalent because the expression is continuous on the set of α on which it is well-defined. This completes the proof. □

C Proof of Theorem 4: SDP tightness

Theorem 4. *For any $K \le D$, any C, and any $\mathbf{w} \in \mathbb{R}^D$, it holds that $\mathcal{R}_{K,C}(\mathbf{w}) = \mathcal{R}_K^{SDP}(\mathbf{w})$.*

The high-level idea of the proof of Theorem 4 is to take an optimal solution $\mathbf{Z}$ to eq. (10), and construct a rank 1 solution that obtains the same objective and satisfies the constraints. We reiterate the SDP formulation for easy reference:

$$\begin{aligned}
\mathcal{R}_K^{\mathrm{SDP}}(\mathbf{w}) = \;& \min_{\mathbf{Z}\succcurlyeq 0} \;\; \langle \mathbf{Z}, \mathbf{I}\rangle \\
& \text{s.t.,} \;\; \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\mathrm{real}}\rangle = 2\,\mathrm{Re}(\widehat{\mathbf{w}}[d]) \\
& \qquad\;\; \forall_{d\in[D]}, \langle \mathbf{Z}, \mathbf{A}_d^{\mathrm{img}}\rangle = 2\,\mathrm{Im}(\widehat{\mathbf{w}}[d]).
\end{aligned} \tag{SDP}$$

We start with the KKT conditions for (10). The D constraints involving $\mathrm{Re}(\mathbf{w})$ correspond to a dual vector $\boldsymbol{\lambda}^{\mathrm{real}} \in \mathbb{R}^D$; the D constraints involving $\mathrm{Im}(\mathbf{w})$ correspond to a dual vector $\boldsymbol{\lambda}^{\mathrm{img}} \in \mathbb{R}^D$. To simplify these conditions, we take $\boldsymbol{\lambda} \in \mathbb{C}^D$ to be $\boldsymbol{\lambda}^{\mathrm{real}} + i\cdot\boldsymbol{\lambda}^{\mathrm{img}}$ (when $\boldsymbol{\lambda}$ is a dual-optimal solution, $\boldsymbol{\lambda}$ is also a Fourier transform of a real vector). The dual variable for the PSD constraint corresponds to a matrix $\mathbf{\Gamma} \succcurlyeq 0$. In this notation, the KKT conditions are primal feasibility, along with the following constraints:

$$\begin{aligned}\mathbf{\Gamma} &= \mathbf{I} - \begin{bmatrix} \mathbf{0}_K & \overline{\mathbf{F}}_K^\top \mathbf{\Lambda} \overline{\mathbf{F}} \\ \mathbf{F}\overline{\mathbf{\Lambda}}\mathbf{F}_K & \mathbf{0}_D \end{bmatrix} \\ \mathbf{\Gamma} &\succcurlyeq 0 \\ \mathbf{Z}\overline{\mathbf{\Gamma}} &= 0.\end{aligned}$$

Now, to simplify these conditions, suppose that $\mathbf{Z}$ is rank L, in which case we can express it as $\mathbf{Z} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{U}^\top & \mathbf{V}^\top \end{bmatrix}$, where $\mathbf{U} \in \mathbb{R}^{K \times L}$ and $\mathbf{V} \in \mathbb{R}^{D \times L}$. Using that $\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ is full rank and through some simple algebraic manipulations, we obtain the following formulation of the KKT conditions:

$$\sum_{l=0}^{L} \widehat{\mathbf{U}}[:, l] \odot \widehat{\mathbf{V}}[:, l] = \widehat{\mathbf{w}} \tag{KKT 1}$$

$$\begin{bmatrix} \mathbf{0}_K & \overline{\mathbf{F}}_K^\top \mathbf{\Lambda} \overline{\mathbf{F}} \\ \mathbf{F}\overline{\mathbf{\Lambda}}\mathbf{F}_K & \mathbf{0}_D \end{bmatrix} \preccurlyeq \mathbf{I}_{D+K} \tag{KKT 2}$$

$$\overline{\widehat{\mathbf{V}}} = \overline{\mathbf{\Lambda}} \widehat{\mathbf{U}} \tag{KKT 3}$$

$$\widehat{\mathbf{U}} = \mathbf{F}_K \overline{\mathbf{F}}_K^\top \mathbf{\Lambda} \overline{\widehat{\mathbf{V}}}. \tag{KKT 4}$$

The KKT conditions give useful properties of the solution $\mathbf{Z}$. For $0 \le l \le L-1$, we let $\mathbf{u}_l = \mathbf{U}[:, l]$ and $\mathbf{v}_l = \mathbf{V}[:, l]$. From (KKT 3), we see that

$$\forall_{l \in [L]} \widehat{\mathbf{v}}_l = \boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}_l}.$$

Combining the above with (KKT 1), we obtain:

$$\widehat{\mathbf{w}} = \sum_{l=0}^{L-1} \boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}_l} \odot \widehat{\mathbf{u}}_l.$$

We now make the following assertion.

Claim. *We claim that in order to prove Theorem 4, it suffices to find a vector* $\mathbf{u} \in \mathbb{R}^K$ *such that:*

$$\overline{\widehat{\mathbf{u}}} \odot \widehat{\mathbf{u}} = \sum_{l=0}^{L-1} \overline{\widehat{\mathbf{u}}_l} \odot \widehat{\mathbf{u}}_l. \tag{CONV-REDUCTION}$$

The most technical part of the proof is to show (CONV-REDUCTION). We will shortly prove the existence of $\mathbf{u}$ in (CONV-REDUCTION) by proving the intermediate Lemma 5. But before that we will first justify our above claim.

Proof of claim. Assume (CONV-REDUCTION) holds and let $\mathbf{u}$ be a solution that satisfies eq. (CONV-REDUCTION), and let $\mathbf{v} = \mathbf{F}^{-1}(\boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}})$ such that we have $\widehat{\mathbf{v}} = \boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}}$. We can take $\mathbf{Z}^*$ to be $\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \begin{bmatrix} \mathbf{u}^\top & \mathbf{v}^\top \end{bmatrix}$.

We can see that $(\mathbf{u}, \mathbf{v})$ satisfies the following:

$$\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}} = \boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}} \odot \widehat{\mathbf{u}} = \sum_{l=0}^{L-1} \boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}_l} \odot \widehat{\mathbf{u}}_l = \widehat{\mathbf{w}}. \tag{19}$$

Thus, $\mathbf{Z}^*$ satisfies the feasibility condition for $\mathcal{R}_K^{\mathrm{SDP}}(\mathbf{w})$. Moreover, we show that the solution also achieves the optimum objective value for $\mathcal{R}_{K,C}(\mathbf{w})$ as follows:

$$\begin{aligned}
\langle \mathbf{Z}^*, \mathbf{I} \rangle &= \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 = \|\widehat{\mathbf{u}}\|_2^2 + \|\widehat{\mathbf{v}}\|_2^2 \\
&= \sum_{d=0}^{D-1} |\widehat{\mathbf{u}}_d|^2 + \sum_{d=0}^{D-1} |\widehat{\mathbf{v}}_d|^2 \\
&\overset{(a)}{=} \sum_{d=0}^{D-1} \sum_{l=0}^{L-1} |(\widehat{\mathbf{u}}_l)_d|^2 + \sum_{d=0}^{D-1} \sum_{l=0}^{L-1} |\boldsymbol{\lambda}_d|^2 |(\widehat{\mathbf{u}}_l)_d|^2 \\
&\overset{(b)}{=} \sum_{l=0}^{L-1} \sum_{d=0}^{D-1} |(\widehat{\mathbf{u}}_l)_d|^2 + \sum_{l=0}^{L-1} \sum_{d=0}^{D-1} |(\widehat{\mathbf{v}}_l)_d|^2 \\
&= \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2 \\
&= \langle \mathbf{Z}, \mathbf{I} \rangle = \mathcal{R}_K^{\mathrm{SDP}}(\mathbf{w})
\end{aligned} \tag{20}$$

where (a) follows from (CONV-REDUCTION) and using $\widehat{\mathbf{v}} = \boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}}$, which follows by definition of $\mathbf{v}$, and (b) follows from (KKT 3) that $\widehat{\mathbf{v}}_l = \boldsymbol{\lambda} \odot \overline{\widehat{\mathbf{u}}}_l$.

Thus, we have show that (CONV-REDUCTION) implies that there exists a rank 1 solution $\mathbf{Z}^*$ that achieves the same objective value as $\mathbf{Z}$ and satisfies the constraints, and hence is minimizer of the SDP as desired. □

Proof of (CONV-REDUCTION)**.** It is more convenient to write (CONV-REDUCTION) in signal space. Taking inverse Fourier transforms of (CONV-REDUCTION), we need to show that given $\mathbf{u}_l = \mathbf{F}_K^\top \widehat{\mathbf{u}}_l \in \mathbb{R}^K$ there exists $\mathbf{u} \in \mathbb{R}^K$ such that the following holds:

$$\mathbf{u} \star \mathbf{u} = \sum_{l=0}^{L-1} \mathbf{u}_l \star \mathbf{u}_l, \tag{21}$$

where the convolutions are taken in D dimensional space, *i.e.,* $\{\mathbf{u}_l\}_l$ are padded with $D - K$ zeros so that $\mathbf{u}_l \star \mathbf{u}_l \in \mathbb{R}^D$.

We can now see that eq. 21 (and hence (CONV-REDUCTION)) indeed holds by recursively applying Lemma 5, which was stated earlier in the main text and is reiterated below.

Lemma 5. *For any $1 \leq K \leq D$, and for any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$, there exists a vector $\mathbf{c} \in \mathbb{R}^K$ such that $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b} = \mathbf{c} \star \mathbf{c}$, where convolutions are w.r.t. dimension D.*

The proof of Lemma 5 is provided in Section C.1.

We have thus far shown that Theorem 4 follows from Lemma 5: we first established that it suffices to find $\mathbf{u} \in \mathbb{R}^K$ satisfying (21) (or equivalently (CONV-REDUCTION)), which in turn holds from recursively applying Lemma 5. Now, it suffices to prove Lemma 5; we prove this lemma in the following subsections.

C.1 Proof of Lemma 5

For $K = D$, Lemma 5 follows easily from the Fourier space representation, since the Fourier space representation of the vector $\mathbf{c}$ can be explicitly constructed as the square root of the Fourier representation of $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b}$. However, this construction does not generalize to kernel sizes $K < D$, and Lemma 5 is thus non-trivial in general. In fact, the vector $\mathbf{c}$ does not even appear to have a clean closed-form characterization for general K. To sidestep this issue, we use a proof technique that enables us to implicitly construct the vector $\mathbf{c}$. We believe that this proof technique could be of independent interest.

C.1.1 Reducing Lemma 5 to $D = 2K - 1$ Case

The first step in the proof of Lemma 5 is to show that Lemma 5 for general K, D follows from the special case where $D = 2K - 1$.

Lemma 14. *For any $K \geq 1$, for $D = 2K - 1$, and for any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$, there exists a vector $\mathbf{c} \in \mathbb{R}^K$ such that $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b} = \mathbf{c} \star \mathbf{c}$, where convolutions are taken with respect to the base dimension D.*

From Lemma 14, it is not difficult to conclude Lemma 5.

Proof of Lemma 5 from Lemma 14. By Lemma 14, we know that Lemma 5 holds when $D = 2K-1$. We now show that this implies the statement for a general value of D. Regardless of D, we can take the coordinates in $[-K+1, K-1]$ of $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b}$ and apply Lemma 14 to obtain a vector $\mathbf{c} \in \mathbb{R}^K$. We now show that $\mathbf{c}$ satisfies $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b} = \mathbf{c} \star \mathbf{c}$, regardless of the value of base dimension D.

Case 1: $D \geq 2K - 1$**.** First, notice that for $D' \notin [-K+1, K-1] \mod D$, we know that

$$(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_d = 0 = (\mathbf{c} \star \mathbf{c})_d,$$

because the vectors are K-dimensional and thus their convolutions will only have at most $2K - 1$ nonzero entries. For $d \in [-K+1, K-1] \mod D$, we see that by Lemma 14:

$$(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_d = (\mathbf{c} \star \mathbf{c})_d.$$

for $d \in [-K+1, K-1] \mod D$.

Case 2: $D < 2K - 1$. By definition, we also know that $K \leq D$. We see that $(\mathbf{c} \star \mathbf{c})_d$ (where the convolution is taken with base dimension D) is equal to the sum over all d' such that d and d' are equal mod D of $(\mathbf{c} \star \mathbf{c})_{d'}$ (where the convolution is taken with base dimension $2K - 1$). Similarly, $(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_d$ (where the convolution is taken with base dimension D) is equal to the sum over all d' such that d and d' are equal mod D of $(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_{d'}$ (where the convolution is taken with base dimension $2K - 1$). By Lemma 14, we know that $(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_d = (\mathbf{c} \star \mathbf{c})_{d'}$ for all d' where the convolution is taken with base dimension $2K - 1$. This means that $(\mathbf{c} \star \mathbf{c})_d = (\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_d$ where the convolution is taken with base dimension D. $\square$

The remainder of the section of devoting to proving Lemma 14. This statement trivially holds with $\mathbf{c}$ as the zero vector if $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b} = 0$, so for the remainder of the proof, we assume that $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b}$ is nonzero.

C.1.2 Introducing the polynomial representation

The key idea for the proof of Lemma 14 is to use the polynomial formulation of convolutions.[9] We then use factorization of the polynomials to implicitly construct $\mathbf{c}$. We use the following notation. Let $\mathcal{P}_k \subseteq \mathbb{R}[x]$ denote the set of degree $\leq K-1$ polynomials with *real coefficients*. For a vector $\mathbf{z} \in \mathbb{R}^K$, we define the *polynomial representation* $p_{\mathbf{z}}(x) \in \mathcal{P}_{K-1}$ to be the polynomial $\mathbf{z}_0 + \mathbf{z}_1 x + \ldots + \mathbf{z}_{K-1}x^{K-1}$. Using the polynomial representations, convolutions can be expressed as polynomial multiplication:

Fact 1. *Let $\mathbf{a} \in \mathbb{R}^K$. The $D = 2K-1$ dimensional convolution $\mathbf{a} \star \mathbf{a}$ has polynomial representation $p_{\mathbf{a}\star\mathbf{a}}(x)$ that is equivalent to the polynomial $x^{K-1}p_{\mathbf{a}}(x) \cdot p_{\mathbf{a}}(1/x) \in \mathcal{P}_{2K-2}$ up to permuting the coefficients appropriately.*

The polynomial representation enables us to construct a vector $\mathbf{c}$ in terms of the roots of the relevant polynomials. We now reformulate Lemma 14 using the polynomial representation. Recall that we wish to show that there exists a vector $\mathbf{c} \in \mathbb{R}^K$ such that

$$\mathbf{c} \star \mathbf{c} = \mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b}.$$

Using the polynomial representation, we can equivalently write this requirement as $p_{\mathbf{c}\star\mathbf{c}}(x) = p_{\mathbf{a}\star\mathbf{a}}(x) + p_{\mathbf{b}\star\mathbf{b}}(x)$. Now, applying Fact 1, we see that the polynomial representation of the lemma statement is the following:

$$x^{K-1}p_{\mathbf{c}}(x) \cdot p_{\mathbf{c}}(1/x) = x^{K-1}p_{\mathbf{a}}(x)p_{\mathbf{a}}(1/x) + x^{K-1}p_{\mathbf{b}}(x)p_{\mathbf{b}}(1/x).$$

To simplify notation, we denote the right-hand-side of the previous equation by $Q(x)$:

$$Q(x) := x^{K-1}p_{\mathbf{a}}(x)p_{\mathbf{a}}(1/x) + x^{K-1}p_{\mathbf{b}}(x)p_{\mathbf{b}}(1/x) \in \mathcal{P}_{2K-2}.$$

In this notation, our goal is to show the following:

$$Q(x) = x^{K-1}p_{\mathbf{c}}(x) \cdot p_{\mathbf{c}}(1/x).$$

Since there is a 1-to-1 correspondence between polynomials in $\mathcal{P}_{K-1}$ and vectors in $\mathbb{R}^K$, it suffices to show that there exists a polynomial $p \in \mathcal{P}_{K-1}$ such that:

$$Q(x) = x^{K-1}p(x) \cdot p(1/x). \tag{22}$$

The remainder of the proof boils down to constructing $p \in \mathcal{P}_{K-1}$ such that eq. (22) is satisfied.

C.1.3 Proving the polynomial representation version of the lemma statement

The first property of $Q(x)$ that we leverage is that it is a *palindromic polynomial* (i.e. a polynomial where the coefficients form a palindrome) with real coefficients. To see that $Q(x)$ is a palindromic polynomial, notice that $(\mathbf{a}\star\mathbf{a})_d = (\mathbf{a}\star\mathbf{a})_{D-d}$ and $(\mathbf{b}\star\mathbf{b})_d = (\mathbf{b}\star\mathbf{b})_{D-d}$ for all $0 \leq d \leq D-1$, and so $(\mathbf{a}\star\mathbf{a}+\mathbf{b}\star\mathbf{b})_d = (\mathbf{a}\star\mathbf{a}+\mathbf{b}\star\mathbf{b})_{D-d}$ for all $0 \leq d \leq D-1$. Now, we can use Fact 1 to conclude that the x^d coefficient of $Q(x)$ is the $(d-K+1) \mod D$ coefficient of $\mathbf{a}\star\mathbf{a}+\mathbf{b}\star\mathbf{b}$. Thus, the x^d coefficient of $Q(x)$ is equal to the x^{2K-2-d} coefficient of $Q(x)$, as desired.

At first glance, it would appear that eq. 22 follows immediately from standard properties of palindromic polynomials with real coefficients whose roots are known to come in reciprocal pairs.[10] However, we cannot obtain eq. (22) from the palindromic property alone. To see this, consider the following example:

[9] E.g., see `https://en.wikipedia.org/wiki/Convolution`

[10] This is a standard fact: e.g., see `https://en.wikipedia.org/wiki/Reciprocal_polynomial`.

Example 1. *Consider the palindromic polynomial with real coefficients $x^2+1 = ix(x-i)(1/x-i)$. This polynomial is not expressible as $xp'(x)p'(1/x)$ for any **real** polynomial p'.*

The proof of eq. (22) thus must leverage further structure of $Q(x)$, which we will ultimately extract through examining the roots of $Q(x)$. Using that $\mathbb{C}$ is algebraically closed, we can factor $Q(x)$ into a polynomial $c_Q \prod_i (x-\alpha_i)$ with exactly $2K-2$ roots where the α_i need not be distinct. To show eq. (22), it suffices to show that $x^{K-1}p(x)\cdot p(1/x)$ has roots (with multiplicities) given by the multi-set $S_Q = \{\alpha_i\}$ and has leading coefficient c. Drawing upon this formulation, we will construct p implicitly by the multi-set S_p of its roots (with multiplicities) and its leading coefficient $c_p \neq 0$: that is, so that $p = c_p \prod_{\alpha \in S_p}(x-\alpha)$.

C.1.4 Property of the roots of $Q(x)$

Before we construct S_p and c_p, it is helpful to establish properties of the multi-set of roots S_Q.

1. (P1) For every root α with multiplicity m, $\overline{\alpha}$ is a root and has multiplicity m.
2. (P2) If $\alpha \neq 0$ is a root with multiplicity m, then $1/\alpha$ is a root with multiplicity m.
3. (P3) If α such that $|\alpha| = 1$ is a root, then α has even multiplicity.

The first two properties (P1) and (P2) follow from the fact that $Q(x)$ is a palindromic polynomial with real coefficients. In particular, we see that (P1) follows from the fact that $Q(x)$ has real coefficients so that the roots come in conjugate pairs, and (P2) follows from standard properties of palindromic polynomials.[11]

Establishing the critical property of $Q(x)$. The last property, (P3), uses deeper aspects of the structure of $Q(x)$. In particular, it uses that $Q(x)$ is the sum of polynomials of the form $x^{K-1}p'(x)p'(1/x)$ where $p' \in \mathcal{P}_{K-1}$ has real coefficients, rather than just an arbitrary palindromic polynomial. To see this, let's return to Example 1 and observe that x^2+1 does not satisfy (P3) since its roots are i and $-i$. Hence, we use further structure of $Q(x)$ and we show:

Lemma 15. *Consider vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$, and let $p_{\mathbf{a}}, p_{\mathbf{b}} \in \mathcal{P}_{K-1}$ be their polynomial representation. If $\alpha \in \mathbb{C}$ such that $|\alpha| = 1$ is a root of $p_{\mathbf{a}}(x)(x^{K-1}p_{\mathbf{a}}(1/x)) + p_{\mathbf{b}}(x)(x^{K-1}p_{\mathbf{b}}(1/x))$, then α has even multiplicity.*

Proof. Suppose that α is a root of $Q(x)$ and $|\alpha| = 1$. We see that

$$
\begin{aligned}
0 &= Q(\alpha) \\
&= p_{\mathbf{a}}(\alpha)(\alpha^{K-1}p_{\mathbf{a}}(1/\alpha)) + p_{\mathbf{b}}(\alpha)(\alpha^{K-1}p_{\mathbf{b}}(1/\alpha)) \\
&= p_{\mathbf{a}}(\alpha)(\alpha^{K-1}p_{\mathbf{a}}(\overline{\alpha})) + p_{\mathbf{b}}(\alpha)(\alpha^{K-1}p_{\mathbf{b}}(\overline{\alpha})) \\
&= \alpha^{K-1}\left(p_{\mathbf{a}}(\alpha)\overline{p_{\mathbf{a}}(\alpha))} + p_{\mathbf{b}}(\alpha)\overline{p_{\mathbf{b}}(\alpha))}\right) \\
&= \alpha^{K-1}\left(|p_{\mathbf{a}}(\alpha)|^2 + |p_{\mathbf{b}}(\alpha)|^2\right).
\end{aligned}
$$

Since $\alpha \neq 0$, this means that $|p_{\mathbf{a}}(\alpha)|^2 + |p_{\mathbf{b}}(\alpha)|^2 = 0$. Thus, $p_{\mathbf{a}}(\alpha) = 0$ and $p_{\mathbf{b}}(\alpha) = 0$. Now, it suffices to show that α is a root with even multiplicity in $p_{\mathbf{a}}(x)(x^{K-1}p_{\mathbf{a}}(1/x))$ and in $p_{\mathbf{b}}(x)(x^{K-1}p_{\mathbf{b}}(1/x))$.

We show that α has even multiplicity in $p_{\mathbf{a}}(x)(x^{K-1}p_{\mathbf{a}}(1/x))$ (an analogous argument shows this for $p_{\mathbf{a}}(x)(x^{K-1}p_{\mathbf{a}}(1/x))$). Suppose that α has multiplicity m in $p_{\mathbf{a}}(x)$. Since $p_{\mathbf{a}}(x)$ has real

[11] This is a standard fact: e.g., see `https://en.wikipedia.org/wiki/Reciprocal_polynomial`.

coefficients, we know that $\overline{\alpha}$ is a root of $p_{\mathbf{a}}(x)$ with multiplicity m. We also know that $1/\alpha = \overline{\alpha}$ is a root with multiplicity m of $x^{k-1}p_{\mathbf{a}}(1/x))$. Since $x^{K-1}p_{\mathbf{a}}(1/x))$ has real coefficients, we know that $\overline{\overline{\alpha}} = \alpha$ is a root with multiplicity m of $x^{k-1}p_{\mathbf{a}}(1/x))$. This means that α has multiplicity $2m$ in $p_{\mathbf{a}}(x)(x^{K-1}p_{\mathbf{a}}(1/x))$ as desired. $\square$

We note that (P3) follows immediately from Lemma 15.

C.1.5 Constructing the roots of p

We construct the multi-set of roots S_p. To do this, we begin by constructing the *nonzero* roots in S_p, and then we add in the zero roots with the appropriate multiplicities at the end.

Constructing the nonzero roots. The high-level intuition for nonzero roots is that we ultimately need the roots of $x^{K-1}p(x)p(1/x)$ to exactly match the roots in S_Q. Intuitively, since the roots of $p(1/x)$ are the reciprocals of the roots of $p(x)$ (with multiplicity preserved), and since polynomials with real coefficients have roots in conjugate pairs, we wish to divide the real roots into disjoint pairs $(\alpha, 1/\alpha)$ (so that $p(x)$ has root α and $p(1/x)$ has root $1/\alpha$) and the complex roots into disjoint quadruples $(\alpha, \overline{\alpha}, 1/\alpha, 1/\overline{\alpha})$ (so that $p(x)$ has roots α and $\overline{\alpha}$ and $p(1/x)$ has roots $1/\alpha$ and $1/\overline{\alpha}$).

Let's formalize this intuition by constructing an undirected graph G where the vertices are the *nonzero* roots in S_Q (a root with multiplicity m corresponds to m separate vertices). The edges are defined as follows. When $\alpha \neq 1$, we connect a vertex corresponding to α with some vertex corresponding to $1/\alpha$, so that the graph forms a bipartite graph (it is possible to do this because of (P2)). By (P3), we can handle $\alpha = 1$, and form non-self-loop edges of the form $(1, 1)$. Let G^{real} be the subgraph of G consisting of vertices corresponding to real roots, and let G^{complex} be the subgraph of G consisting of vertices corresponding to roots with nonzero imaginary part. It is easy to see that $G = G^{\text{real}} \cup G^{\text{complex}}$ and these graphs are disjoint.

Let us now use these graphs to construct the set of nonzero roots in S_p, which will contain half of the vertices in G. For G^{real}, we add one vertex from each edge in G^{real} to S'_p. For G^{complex}, we can pair up the edges in the graph G^{complex} as follows. When $|\alpha| \neq 1$, we can pair up the edges $(\alpha, 1/\alpha)$ and $(\overline{\alpha}, 1/\overline{\alpha})$ so that the pairs are disjoint and no edge is paired with itself (it is possible to do this because of (P1), coupled with the fact that $1/\alpha \neq \overline{\alpha}$). When $|\alpha| = 1$, we use the fact that α has even multiplicity (see (P3)). Thus, we can pair up each edge $(\alpha, 1/\alpha)$ with an edge of the form $(\alpha, 1/\alpha) = (\overline{\alpha}, 1/\overline{\alpha})$, so that pairs continue to be disjoint and no edge is paired with itself. Now, for each pair $(\alpha, 1/\alpha)$ and $(\overline{\alpha}, 1/\overline{\alpha})$, we add α to S_p and we add $\overline{\alpha}$ to S_p.

Adding the zero roots. To construct S_p, all that remains is to determine the multiplicity of the zero roots. Let m be the multiplicity of 0 in S_Q. We then simply add m copies of 0 to S_p.

C.1.6 Proof that S_p is the correct multi-set of roots

We first prove that S_p corresponds to the roots of a degree $\leq K-1$ polynomial with real coefficients; we then show that the multiset of roots of $x^{K-1}p(x)p(1/x)$ is equal to S_Q.

Proof that S_p corresponds to the roots of a polynomial in $\mathcal{P}_{K-1}$. For S_p to be a valid multi-set of roots for a polynomial $p \in \mathcal{P}_{K-1}$, we need to ensure that S_p consists of at most $K-1$ elements and that the roots come in conjugate pairs.

To see that S_p consists of at most $K-1$ elements, we do the following root counting argument. Notice that if $(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_d = 0$, this means that $(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})_{D-d} = 0$. In terms of the polynomial representation, this means that if $Q(x)$ has a zero root with multiplicity m, then the degree of $Q(x)$ is $2K-2-m$ (since $Q(x) \neq 0$). Thus, we see that $Q(x)$ has $2K-2-m$ roots not including multiplicity, and $2K-2-2m$ nonzero roots. By the construction of S, we see that $p_{\text{monic}}(x)$ has $K-1-m$ nonzero roots including multiplicities, and m zero roots, which amounts to $K-1$ roots in total with multiplicity, as desired.

The fact that the roots come in conjugate pairs follows from the construction of the graphs above. Recall that the real roots of S_Q were partitioned into disjoint pairs $(\alpha, 1/\alpha)$ (so that $p(x)$ has root α) and the complex roots of S_Q were partitioned into disjoint quadruples $(\alpha, \overline{\alpha}, 1/\alpha, 1/\overline{\alpha})$ (so that $p(x)$ has roots α and $\overline{\alpha}$). This ensures that the roots comes up in conjugate pairs as desired.

Proof that the multi-set of roots of $x^{K-1}p(x)p(1/x)$ equals the multi-set S_Q. Let $p_{\text{monic}}(x)$ be the monic polynomial given by the roots of S_p, and consider

$$Q_1(x) := p_{\text{monic}}(x)x^{K-1}p_{\text{monic}}(1/x).$$

It suffices to show that the multi-set of roots of $Q_1(x)$ with multiplicities is equal to the multi-set S_Q.

For the nonzero roots, we use the construction of S_p. Recall that the real roots of S_Q were partitioned into disjoint pairs $(\alpha, 1/\alpha)$ (so that $p(x)$ has root α) and the complex roots of S_Q were partitioned into disjoint quadruples $(\alpha, \overline{\alpha}, 1/\alpha, 1/\overline{\alpha})$ (so that $p(x)$ has roots α and $\overline{\alpha}$). This, coupled with the fact that the nonzero roots of $p(1/x)$ are the inverses of the nonzero roots of $p(x)$, means that the union of the multi-set of nonzero roots in $p(x)$ with the multi-set of nonzero roots of $p(1/x)$ is equal to the multi-set of nonzero roots of $Q(x)$. This means that the multi-set of roots of $Q_1(x)$ with multiplicities is equal to the multi-set S_Q.

For the zero roots, we simply need to show that $x^{K-1}p_{\text{monic}}(1/x)$ has no roots that are 0. Using that the constant term of $x^{K-1}p_{\text{monic}}(1/x)$ is equal to the coefficient of x^{K-1} in $p_{\text{monic}}(x)$, it suffices to show that the coefficient of x^{K-1} in $p_{\text{monic}}(x)$ is nonzero. To see this, we use the fact that $p_{\text{monic}}(x)$ has $K-1$ roots in total with multiplicity by the argument from the previous paragraph, and so the degree of $p_{\text{monic}}(x)$ must be $K-1$. This means that the roots of $Q_1(x)$ with multiplicities are equal to the roots of $Q(x)$.

C.1.7 Handling the leading coefficient of p

Now, we need to just construct the leading coefficient of p. As above, let $p_{\text{monic}}(x) = \prod_{\alpha \in S_p}(x-\alpha)$ be the monic polynomial given by the roots of S_p, and consider $Q_1(x) = p_{\text{monic}}(x)x^{K-1}p_{\text{monic}}(1/x)$. Since $Q_1(x)$ and $Q(x)$ have the same set of roots with multiplicities, we know that $Q_1(x) = \gamma \cdot Q(x)$ for some $\gamma \neq 0$. Let's take $c_p = \sqrt{\gamma}$. In order for $p = c_p \prod_{\alpha \in S_p}(x-\alpha)$ to be a polynomial with real coefficients, we need c_p to be real.

To show that c_p is real, it suffices to show that γ is positive. This can be seen as follows. Let $\mathbf{c}_{\text{monic}} \in \mathbb{R}^D$ be the vector with polynomial representation $p_{\text{monic}}(x)$. Now, by Fact 1, we know that the polynomial representation $p_{\mathbf{c}_{\text{monic}} \star \mathbf{c}_{\text{monic}}}$ takes the form $x^D p_{\text{monic}}(x) p_{\text{monic}}(1/x) = x^{D-K+1}Q_1(x)$, where exponents are taken modulo D. Since $Q_1(x) = \gamma \cdot Q(x)$, and since the polynomial representation of $(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})$ takes the form $x^{D-K+1}Q(x)$ where the exponents are taken modulo D, we can conclude that $\mathbf{c}_{\text{monic}} \star \mathbf{c}_{\text{monic}} = \gamma(\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b})$, and thus that the Fourier representations of these vectors are also equal. It is not difficult to verify that the Fourier representation of $\mathbf{c}_{\text{monic}} \star \mathbf{c}_{\text{monic}}$ is $|\widehat{\mathbf{c}}_{\text{monic}}|^2$, and so the entries of the Fourier representation of consist of all nonnegative real numbers. Similarly, Fourier representation of $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b}$ is $|\widehat{\mathbf{a}}|^2 + |\widehat{\mathbf{b}}|^2$ which consists of entries that are all nonnegative real numbers. This means that $\gamma \geq 0$, as desired.

C.1.8 Concluding Lemma 14

We have shown that $p = c_p \prod_{\alpha \in S_p}(x - \alpha)$ is a polynomial of degree at most $K - 1$ with real coefficients. Moreover, we have shown that $Q(x)$ has the same multi-set of roots and the same leading coefficient as $x^{K-1}p(x) \cdot p(1/x)$. Thus, we can conclude that eq. (22) holds, and thus we have proven the polynomial representation of Lemma 14. This concludes the proof of Lemma 14.

C.2 Discussion of the proof technique

We conclude with a discussion of the analysis and highlight the main parts of the proof. At the beginning of the section, we used the KKT conditions to show that it suffices to prove CONV-REDUCTION, an additive property about convolutions of for kernel size. We then showed that it suffices to prove a version of this statement for the sum of two such convolutions, i.e. Lemma 5. We believe that this property could be of independent interest.

The bulk of the proof boiled down to proving Lemma 5 in the special case of $D = 2K - 1$, i.e. Lemma 14. Proving Lemma 14 was the core technical contribution in this section. Since $\mathbf{c}$ does not necessarily always a clean closed-form solution as a function of $\mathbf{a}$ and $\mathbf{b}$, we needed to construct $\mathbf{c}$ *implicitly.* The polynomial representation of convolutions enabled us to implicitly construct $\mathbf{c}$ via its roots. To construct $\mathbf{c}$ and ensure that the corresponding polynomial representation p had real coefficients, we needed to leverage the structure of the polynomial representation $Q(x)$ of $\mathbf{a} \star \mathbf{a} + \mathbf{b} \star \mathbf{b}$ beyond its palindromic structure. (This additional property was proven in Lemma 15.) With this structure, we were able to factor $Q(x)$ and partition its roots in order to construct the roots of p.

D Remaining proofs of results in Section 4

D.1 Proof of Corollary 6

Corollary 6. *For $K \leq D$ and any C, $\mathcal{R}_{K,C}(\mathbf{w})$ is a norm.*

Corollary 6 follows from Theorem 4.

Proof of Corollary 6. It suffices to establish the scalar multiplication property, the triangle inequality, and point separation.

Scalar multiplication. Let $\gamma \in \mathbb{R}$. By definition, we see that

$$\mathcal{R}_{K,C}(\gamma \mathbf{w}) = \min_{\mathbf{U},\mathbf{V}} \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2 \quad \text{s.t.,} \quad \operatorname{diag}(\widehat{\mathbf{U}}\widehat{\mathbf{V}}^\top) = \gamma \widehat{\mathbf{w}}.$$

Let's do a change of variables $\mathbf{U} \leftarrow \frac{1}{\sqrt{|\gamma|}}\mathbf{U}, \mathbf{V} \leftarrow \frac{1}{\operatorname{sign}(\gamma)\sqrt{|\gamma|}}\mathbf{V}$ to see that

$$\mathcal{R}_{K,C}(\gamma \mathbf{w}) = |\gamma| \left[\min_{\mathbf{U},\mathbf{V}} \|\mathbf{U}\|^2 + \|\mathbf{V}\|^2 \quad \text{s.t.,} \quad \operatorname{diag}(\widehat{\mathbf{U}}\widehat{\mathbf{V}}^\top) = \widehat{\mathbf{w}} \right] = |\gamma| \mathcal{R}_{K,C}(\mathbf{w})$$

as desired.

Triangle inequality. It suffices to show that if $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2$, then $\mathcal{R}_{K,C}(\mathbf{w}) \leq \mathcal{R}_{K,C}(\mathbf{w}_1) + \mathcal{R}_{K,C}(\mathbf{w}_2)$. By Theorem 4, it suffices to show that $\mathcal{R}_K^{\text{SDP}}(\mathbf{w}) \leq \mathcal{R}_{K,C}(\mathbf{w}_1) + \mathcal{R}_{K,C}(\mathbf{w}_2)$. Suppose that $\mathbf{U}_1$ and $\mathbf{V}_1$ are optimal solutions to $\mathcal{R}_{K,C}(\mathbf{w}_1)$, and suppose that $\mathbf{U}_2$ and $\mathbf{V}_2$ are optimal solutions to $\mathcal{R}_{K,C}(\mathbf{w}_2)$. If we define:

$$\begin{aligned}
\mathbf{Z}_1 &= \begin{bmatrix} \mathbf{U}_1\mathbf{U}_1^\top & \mathbf{U}_1\mathbf{V}_1^\top \\ \mathbf{V}_1\mathbf{U}_1^\top & \mathbf{V}_1\mathbf{V}_1^\top \end{bmatrix} \\
\mathbf{Z}_2 &= \begin{bmatrix} \mathbf{U}_2\mathbf{U}_2^\top & \mathbf{U}_2\mathbf{V}_2^\top \\ \mathbf{V}_2\mathbf{U}_2^\top & \mathbf{V}_2\mathbf{V}_2^\top \end{bmatrix} \\
\mathbf{Z} &= \mathbf{Z}_1 + \mathbf{Z}_2,
\end{aligned}$$

then we see that the SDP objective $\langle \mathbf{Z}, \mathbf{I} \rangle = \langle \mathbf{Z}_1, \mathbf{I} \rangle + \langle \mathbf{Z}_2, \mathbf{I} \rangle = \mathcal{R}_{K,C}(\mathbf{w}_1) + \mathcal{R}_{K,C}(\mathbf{w}_2)$. Moreover, we see that $\mathbf{Z}$ is a feasible solution to (10) for $\mathbf{w}$. This means that $\mathcal{R}_k^{\text{SDP}}(\mathbf{w}) \leq \mathcal{R}_{K,C}(\mathbf{w}_1) + \mathcal{R}_{K,C}(\mathbf{w}_2)$ as desired.

Point separation. Notice that $\mathcal{R}_{K,C}(\mathbf{w}) = 0$, then there exist $\mathbf{U}$ and $\mathbf{V}$ such that $\|\mathbf{U}\|^2 + \|\mathbf{V}\|^2 = 0$. This means that $\mathbf{U} = 0$ and $\mathbf{V} = 0$, which means that $\mathbf{w} = 0$ as desired. Moreover, if $\mathbf{w} = 0$, then it's clear that $\mathcal{R}_{K,C}(\mathbf{w}) = 0$. □

D.2 The dual formulation of SDP

In order to analyze the SDP formulation, we consider the dual. We use the formulation of the dual variable in C as $\boldsymbol{\lambda} \in \mathbb{C}^D$. In this form, the dual can be expressed as:

$$\begin{aligned}
\max_{\boldsymbol{\lambda} \in \mathbb{C}^D} \quad & \operatorname{Re}\left(\langle \boldsymbol{\lambda}, \widehat{\mathbf{w}} \rangle\right) \\
\text{s.t.} \quad & \begin{bmatrix} \mathbf{0}_K & \mathbf{F}_K^\top \overline{\boldsymbol{\Lambda}} \mathbf{F} \\ \overline{\mathbf{F}} \boldsymbol{\Lambda} \overline{\mathbf{F}}_K & \mathbf{0}_D \end{bmatrix} \preccurlyeq \mathbf{I}.
\end{aligned}$$

To simplify the objective $\operatorname{Re}\left(\langle \boldsymbol{\lambda}, \widehat{\mathbf{w}} \rangle\right)$, notice that the phases of $\boldsymbol{\lambda}$ can be set to align with $\widehat{\mathbf{w}}$ without affecting the constraint. Thus, we can set the objective to be $|\langle \boldsymbol{\lambda}, \widehat{\mathbf{w}} \rangle|$. For convenience, we also expand out the conic constraint in vector form, and this reformulation incurs a factor of 2 on the objective. We thus obtain the following equivalent formulation of the dual:

$$\begin{aligned}
\max_{\boldsymbol{\lambda} \in \mathbb{C}^D} \quad & 2 \sum_{d=0}^{D-1} |\boldsymbol{\lambda}[d]| \cdot |\mathbf{w}[d]| \\
\text{s.t.} \quad & \forall \mathbf{x} \in \mathbb{C}^K, \quad \sum_{d=0}^{D-1} |\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 \leq 1.
\end{aligned} \tag{23}$$

We now show that strong duality holds for this SDP.

Proposition 16. *The SDP in* (10) *satisfies strong duality.*

Proof. To show strong duality, it suffices to show Slater's condition. We just need to find a solution $\boldsymbol{\lambda} \in \mathbb{C}^d$ where the inequality constraint is not tight. That is, we need to find $\boldsymbol{\lambda}$ such that $\forall_{\mathbf{x} \in \mathbb{C}^K}$, $\sum_{d=0}^{D-1} |\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 < 1$. Let's take $\boldsymbol{\lambda} = [1/2, 0, 0, \ldots, 0]$. Notice that $\sum_{d=0}^{D-1} |\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 = 0.5 |\widehat{\mathbf{x}}[0]|^2 \leq 0.5 < 1$, as desired. □

With the dual, along the fact that $\mathcal{R}_K^{\text{SDP}}(\mathbf{w})$ is a norm, we are equipped to prove general upper and lower bounds on the induced regularizer as well as sharper bounds for patterned vectors.

D.3 Proof of Lemma 7

Lemma 7. *For any $K \le D$, any C, and any $\mathbf{w} \in \mathbb{R}^D$:*

$$2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_2 \le \mathcal{R}_{K,C}(\mathbf{w}) \le 2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$$
$$2\|\widehat{\mathbf{w}}\|_1 \le \mathcal{R}_{K,C}(\mathbf{w}) \le 2\sqrt{\left\lceil\frac{D}{K}\right\rceil}\|\widehat{\mathbf{w}}\|_1.$$

Proof of Lemma 7. The bounds of $2\|\widehat{\mathbf{w}}\|_1$ and $2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$ follow in a straightforward way from Lemma 1—2 coupled with Theorem 4.

The lower bound of $2\|\widehat{\mathbf{w}}\|_1$ follows as: $2\|\widehat{\mathbf{w}}\|_1 \overset{(a)}{=} \mathcal{R}_{D,1}(\mathbf{w}) \overset{(b)}{=} \mathcal{R}_{D,C}(\mathbf{w}) \overset{(c)}{\le} \mathcal{R}_{K,C}(\mathbf{w})$, where (a) follows from Lemma 2, (b) follows from Theorem 4, and (c) follows from Remark 1. Similarly, the upper bound of $2\sqrt{D}\|\widehat{\mathbf{w}}\|_2$ follows as: $2\sqrt{D}\|\widehat{\mathbf{w}}\|_2 \overset{(a)}{=} \mathcal{R}_{1,1}(\mathbf{w}) \overset{(b)}{=} \mathcal{R}_{1,C}(\mathbf{w}) \ge \mathcal{R}_{K,C}(\mathbf{w})$, where (a) follows from Lemma 2 and (b) follows from Theorem 4, and (c) follows from Remark 1

The bulk of the proof lies in showing the lower bound of $2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_2$, and an upper bound of $2\sqrt{\lceil\frac{D}{K}\rceil}\|\widehat{\mathbf{w}}\|_1$. We first prove the lower bound, and then we prove the upper bound.

Proof of the lower bound $2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_2$. We prove that $\mathcal{R}_{K,C}(\mathbf{w}) \ge 2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_2$. It suffices to consider a dual feasible vector to eq. (23) that achieves an objective $2\sqrt{\frac{D}{K}}\|\mathbf{w}\|_2$. We consider

$$\boldsymbol{\lambda} = \frac{\mathbf{w}}{\|\mathbf{w}\|}\sqrt{\frac{D}{K}}.$$

We see that the objective is equal to

$$2\sum_{d=0}^{D-1}|\boldsymbol{\lambda}[d]||\mathbf{w}[d]| = 2\sqrt{\frac{D}{K}}\|\mathbf{w}\|_2,$$

as desired. It thus suffices to show that $\boldsymbol{\lambda}$ satisfies $\sum_{d=0}^{D-1}|\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 \le 1$ for all $\mathbf{x} \in \mathbb{C}^d$ such that $\|\mathbf{x}\|_2 \le 1$. Using Holder's inequality, we can bound:

$$\sum_{d=0}^{D-1}|\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 \le \left(\max_{0\le d\le D-1}|\widehat{\mathbf{x}}[d]|^2\right)\left(\sum_{d=0}^{D-1}|\boldsymbol{\lambda}[d]|^2\right) = \left(\max_{0\le d\le D-1}|\widehat{\mathbf{x}}[d]|^2\right)\|\boldsymbol{\lambda}\|_2^2.$$

We can bound the first term by:

$$|\widehat{\mathbf{x}}[d]| = \frac{1}{\sqrt{D}}\left|\sum_{k=0}^{K-1}\mathbf{x}[k]\mathrm{e}^{-2\pi ikd/D}\right| \le \frac{1}{\sqrt{D}}\sum_{k=0}^{K-1}|\mathbf{x}[k]\mathrm{e}^{-2\pi ikd/D}| = \frac{\|x\|_1}{\sqrt{D}} \le \frac{\sqrt{K}}{\sqrt{D}}.$$

Moreover, we see that $\|\boldsymbol{\lambda}\| = \sqrt{\frac{D}{K}}$. This means that $\left(\max_{0\le d\le D-1}|\widehat{\mathbf{x}}[d]|^2\right)\|\boldsymbol{\lambda}\|_2^2 \le 1$, as desired.

Proof of the upper bound $2\sqrt{\lceil \frac{D}{K} \rceil}\|\widehat{\mathbf{w}}\|_1$. We prove that $\mathcal{R}_{K,C}(\mathbf{w}) \leq 2\sqrt{\lceil \frac{D}{K} \rceil}\|\widehat{\mathbf{w}}\|_1$. Our main ingredient is Corollary 6 which tells us that $\mathcal{R}_{K,C}(\mathbf{w})$ is a norm. We define $T = \lceil D/K \rceil$ vectors $\mathbf{w}_0, \ldots, \mathbf{w}_{T-1} \in \mathbb{R}^D$ where $\mathbf{w} = \sum_{t=0}^{T-1} \mathbf{w}_t$, and apply Corollary 6 to obtain that:

$$\mathcal{R}_{K,C}(\mathbf{w}) \leq \sum_{t=0}^{T-1} \mathcal{R}_{K,C}(\mathbf{w}_t).$$

These vectors are chosen that each $\mathcal{R}_{K,C}(\mathbf{w}_t)$ takes on a simple closed-form solution.

In order to construct the vectors $\mathbf{w}_t$, we consider $\widehat{\mathbf{q}} = \sqrt{\widehat{\mathbf{w}}}$, defined so that $\mathbf{w} = \mathbf{q}^{\downarrow} \star \mathbf{q}$ and $\|\widehat{\mathbf{q}}\|^2 = \|\mathbf{q}\|^2 = \|\widehat{\mathbf{w}}\|_1$. We define vectors $\mathbf{r}_0, \ldots, \mathbf{r}_{T-1} \in \mathbb{R}^D$ such that $\sum_{t=0}^{T-1} \mathbf{r}_t = \mathbf{q}$ as follows. Roughly speaking these vectors consist of the disjoint subsets of the coordinates of $\mathbf{q}$ corresponding to the t^{th} block of size K. More formally, for $0 \leq t \leq T-1$, let $\mathbf{r}_t$ be defined so that $\mathbf{r}_t[l] = \mathbf{q}[l]$ for $l \in [t \cdot K, \min((t+1) \cdot K - 1, D-1)]$, and $\mathbf{r}_t[l] = 0$ otherwise. Let $\mathbf{w}_t = \mathbf{r}_t^{\downarrow} \star \mathbf{q}$ for $0 \leq t \leq T-1$. It is evident that $\sum_t \mathbf{r}_t = \mathbf{q}_t$ and hence $\sum_t \mathbf{w}_t = \mathbf{w}$.

We now show that $\mathcal{R}_{K,C}(\mathbf{w}_t) \leq 2\|\mathbf{r}_t\|\|\mathbf{q}\|$. We show this by explicitly constructing solutions to (6), taking advantage of the fact that $\mathbf{r}_t$ is effectively a vector in $\mathbb{R}^K$ that is zero-padded appropriately. This K-dimensional vector $\mathbf{r}'_t \in \mathbb{R}^K$ is given by $\mathbf{r}'_t[k] = \mathbf{r}_t[(t \cdot K + k) \mod D]$ for $0 \leq k \leq K-1$. Now, we wish to write $\mathbf{w}_t$ as a convolution $(\mathbf{r}'_t)^{\downarrow} \star \mathbf{q}_t$, for some suitably chosen vector $\mathbf{q}_t$. Since $\mathbf{w}_t = \mathbf{r}_t^{\downarrow} \star \mathbf{q}$ and $\mathbf{r}'_t$ is merely a circular shifted version of $\mathbf{r}_t$, we can take $\mathbf{q}_t \in \mathbb{R}^D$ to be $\mathbf{q}$ with the coordinates shifted appropriately. Now, we can rescale $\mathbf{r}'_t$ and $\mathbf{q}_t$ so that they have equal ℓ_2 norms, and obtain the following vectors: $\left(\sqrt{\frac{\|\mathbf{q}\|_t}{\|\mathbf{r}'_t\|}}\right) \mathbf{r}'_t$ and $\left(\sqrt{\frac{\|\mathbf{r}'_t\|}{\|\mathbf{q}\|_t}}\right) \mathbf{q}_t$. These vectors are a feasible solution to eq. (6) for $\mathcal{R}_{K,C}(\mathbf{w}_t)$ and achieve an objective of $2\|\mathbf{r}'_t\|\|\mathbf{q}_t\| = 2\|\mathbf{r}_t\|\|\mathbf{q}\|$, as desired.

Using that $\mathcal{R}_{K,C}(\mathbf{w}_t) \leq 2\|\mathbf{r}_t\|\|\mathbf{q}\|$ for $0 \leq t \leq T-1$, we obtain the following bound on $\mathcal{R}_{K,C}(\mathbf{w})$:

$$\mathcal{R}_{K,C}(\mathbf{w}) \leq 2\|\mathbf{q}\| \sum_{t=0}^{T-1} \|\mathbf{r}_t\|.$$

Now, notice that since the supports of $\mathbf{r}_t$ for $0 \leq t \leq T-1$ are disjoint, $\sum_{t=0}^{T-1} \|\mathbf{r}_t\|^2 = \|\mathbf{q}\|^2$. Applying AM-GM, this means that

$$(\sum_{t=0}^{T-1} \|\mathbf{r}_t\|)^2 \leq T(\sum_{t=0}^{T-1} \|\mathbf{r}_t\|^2) = T\|\mathbf{q}\|^2.$$

Thus, we have that

$$\mathcal{R}_{K,C}(\mathbf{w}) \leq 2\sqrt{T}\|\mathbf{q}\|^2 = 2\sqrt{T}\|\widehat{\mathbf{w}}\|_1 = 2\sqrt{\left\lceil \frac{D}{K} \right\rceil}\|\widehat{\mathbf{w}}\|_1.$$

□

D.4 Proof of Lemma 8

Lemma 8. *Consider vectors $\mathbf{w}(\mathbf{p}) = [\mathbf{p}, \mathbf{p}, \ldots, \mathbf{p}] \in \mathbb{R}^D$ specified by $\mathbf{p} \in \mathbb{R}^P$ s.t., P divides D.*
(a) For any $K \leq P$, it holds that $\forall C$:

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) = \frac{D}{P} \cdot \mathcal{R}^{(P)}_{K,1}(\mathbf{p}).$$

(b) For $P \leq K \leq D$ if $K = P \cdot T$ for integer T, then $\forall C$:

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) = 2\frac{D}{\sqrt{TP}}\|\widehat{\mathbf{p}}\|_1 = 2\sqrt{\frac{D}{K}}\|\widehat{\mathbf{w}}\|_1.$$

We first prove an upper bound on $\mathcal{R}_{K,C}(\mathbf{w})$, and then we prove a matching lower bound. In these proofs, let $\mathbf{w} = \mathbf{w}(\mathbf{p})$. Both of these proofs use the standard fact that $\widehat{\mathbf{w}}[(D/P)\cdot p] = \sqrt{\frac{D}{P}}\widehat{\mathbf{p}}[p]$ for $0 \leq p \leq P-1$, and $\widehat{\mathbf{w}}[d] = 0$ if $(D/P) \nmid d$.

Lemma 17 (Upper bound)**.** *Consider vectors $\mathbf{w}(\mathbf{p}) = [\mathbf{p}, \mathbf{p}, \ldots, \mathbf{p}] \in \mathbb{R}^D$ specified by $\mathbf{p} \in \mathbb{R}^P$ (such that D is a multiple of P).*
(a) For any $K \leq P$, it holds that $\forall C$:

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) \leq \frac{D}{P} \cdot \mathcal{R}^{(P)}_{K,1}(\mathbf{p}).$$

(b) For $P \leq K \leq D$ if $K = P \cdot T$ for integer T, then $\forall C$:

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) \leq 2\frac{D}{\sqrt{TP}}\|\widehat{\mathbf{p}}\|_1.$$

Proof of Lemma 17. It suffices to show an upper bound for the case of a single output channel. We explicitly construct a pair $(\mathbf{u}, \mathbf{v})$ where $\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$ achieves the desired objective.

Case 1: $\mathbf{K} \leq \mathbf{P}$. We construct $(\mathbf{u}, \mathbf{v})$ using an optimal solution $\mathbf{u}_\mathbf{p}$ and $\mathbf{v}_\mathbf{p}$ to eq. (6) for $\mathcal{R}^{(P)}_{K,1}(\mathbf{p})$. We let $\mathbf{u}$ be defined so that $\mathbf{u}[k] = \sqrt{\frac{D}{P}}\mathbf{u}_\mathbf{p}[k]$ for $0 \leq k \leq K-1$. We let $\mathbf{v}$ be defined to be $\mathbf{v} = [\mathbf{v}_\mathbf{p}, \ldots, \mathbf{v}_\mathbf{p}]$. Notice that:

$$\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 = \frac{D}{P}\|\mathbf{u}_\mathbf{p}\|^2 + \frac{D}{P}\|\mathbf{v}_\mathbf{p}\|^2 = \frac{D}{P}\mathcal{R}^{(P)}_{K,1}(\mathbf{p}),$$

as desired. To see that $\mathbf{u}^{\downarrow} \star \mathbf{v} = \mathbf{w}$, it suffices to show $\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}} = \mathbf{w}$. Notice that $\widehat{\mathbf{v}}[(D/P)\cdot p] = \frac{\sqrt{D}}{\sqrt{P}}\widehat{\mathbf{v}_\mathbf{p}}[p]$ for $0 \leq p \leq P-1$, and $\widehat{\mathbf{v}}[d] = 0$ if $(D/P) \nmid d$. This means that $(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[d] = 0 = \widehat{\mathbf{w}}[d]$ if $(D/P) \nmid d$ as desired. Thus it suffices to handle $(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[(D/P)\cdot p]$. Notice that $\widehat{\mathbf{v}}[(D/P)\cdot p] = \frac{\sqrt{D}}{\sqrt{P}}\widehat{\mathbf{v}}_\mathbf{p}[p]$, and $\widehat{\mathbf{u}}[(D/P)\cdot p] = \frac{\sqrt{P}}{\sqrt{D}}\frac{\sqrt{D}}{\sqrt{P}}\widehat{\mathbf{u}}_\mathbf{p}[p] = \widehat{\mathbf{u}}_\mathbf{p}[p]$. This means that:

$$(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[(D/P)\cdot p] = \frac{\sqrt{D}}{\sqrt{P}}\widehat{\mathbf{u}}_\mathbf{p}[p]\widehat{\mathbf{v}}_\mathbf{p}[p] = \frac{\sqrt{D}}{\sqrt{P}}\widehat{\mathbf{p}}[p] = \widehat{\mathbf{w}}[(D/P)\cdot p],$$

as desired.

Case 2: $\mathbf{K} = \mathbf{T} \cdot \mathbf{P}$. We construct $(\mathbf{u}, \mathbf{v})$ using an optimal solution $\mathbf{u}_\mathbf{p}$ and $\mathbf{v}_\mathbf{p}$ to eq. (6) for $\mathcal{R}^{(P)}_{P,1}(\mathbf{p})$. We let $\mathbf{u} = \frac{\sqrt{D}}{T^{3/4}\sqrt{P}}[\mathbf{u}_\mathbf{p}, \ldots, \mathbf{u}_\mathbf{p}]$ be a scaled version of T repeated copies of $\mathbf{u}_\mathbf{p}$. We let $\mathbf{v} = \frac{1}{T^{1/4}}[\mathbf{v}_\mathbf{p}, \ldots, \mathbf{v}_\mathbf{p}]$ be a scaled version of $\frac{D}{P}$ repeated copies of $\mathbf{v}_\mathbf{p}$. Notice that

$$\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 = \frac{D}{\sqrt{TP}}\|\mathbf{u}_\mathbf{p}\|^2 + \frac{D}{\sqrt{TP}}\|\mathbf{v}_\mathbf{p}\|^2 = \frac{D}{\sqrt{TP}}\mathcal{R}_{P,1}(\mathbf{p}) = 2\frac{D}{\sqrt{TP}}\|\widehat{\mathbf{p}}\|_1,$$

as desired. To see that $\mathbf{u}^{\downarrow} \star \mathbf{v} = \mathbf{w}$, it suffices to show $\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}} = \widehat{\mathbf{w}}$. Notice that $\widehat{\mathbf{v}}[(D/P)\cdot p] = \frac{\sqrt{D}}{T^{1/4}\sqrt{P}}\widehat{\mathbf{v}}_\mathbf{p}[p]$ for $0 \leq p \leq P-1$, and $\widehat{\mathbf{v}}[d] = 0$ if $(D/P) \nmid d$. This means that $(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[d] = 0 = \widehat{\mathbf{w}}[d]$

if $(D/P) \nmid d$ as desired. Thus it suffices to handle $(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[(D/P) \cdot p]$. Notice that $\widehat{\mathbf{v}}[(D/P) \cdot p] = \frac{\sqrt{D}}{T^{1/4}\sqrt{P}} \widehat{\mathbf{v}}_{\mathbf{p}}[p]$, and $\widehat{\mathbf{u}}[(D/P) \cdot p] = \frac{T\sqrt{P}}{\sqrt{D}} \frac{\sqrt{D}}{\sqrt{P}T^{3/4}} \widehat{\mathbf{u}}_{\mathbf{p}}[p] = T^{1/4} \widehat{\mathbf{u}}_{\mathbf{p}}[p]$. This means that:

$$(\widehat{\mathbf{u}} \odot \widehat{\mathbf{v}})[(D/P) \cdot p] = \frac{\sqrt{D}}{\sqrt{P}} \widehat{\mathbf{u}}_{\mathbf{p}}[p] \widehat{\mathbf{v}}_{\mathbf{p}}[p] = \frac{\sqrt{D}}{\sqrt{P}} \widehat{\mathbf{p}}[p] = \widehat{\mathbf{w}}[(D/P) \cdot p],$$

as desired. □

Lemma 18 (Lower bound)**.** *Consider vectors* $\mathbf{w}(\mathbf{p}) = [\mathbf{p}, \mathbf{p}, \ldots, \mathbf{p}] \in \mathbb{R}^D$ *specified by* $\mathbf{p} \in \mathbb{R}^P$ *(such that* D *is a multiple of* P*).*
(a) For any $K \leq P$*, it holds that* $\forall C$*:*

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) \geq \frac{D}{P} \cdot \mathcal{R}^{(P)}_{K,1}(\mathbf{p}).$$

(b) For $P \leq K \leq D$ *if* $K = P \cdot T$ *for integer* T*, then* $\forall C$*:*

$$\mathcal{R}_{K,C}(\mathbf{w}(\mathbf{p})) \geq 2\sqrt{\frac{D}{K}} \|\widehat{\mathbf{w}}\|_1.$$

Proof of Lemma 18. We use the dual formulation in eq. (23). Our approach is to construct a dual vector $\boldsymbol{\lambda} \in \mathbb{C}^D$ so that eq. (23) achieves the desired objective.

Case 1: $\mathbf{K} \leq \mathbf{P}$. Let $\boldsymbol{\lambda}_{\mathbf{p}} \in \mathbb{C}^P$ be the dual optimal solution for $\mathcal{R}^{\mathrm{SDP}}_K(\mathbf{p})$. Now, let $\boldsymbol{\lambda}[(D/P) \cdot p] = \frac{\sqrt{D}}{\sqrt{P}} \boldsymbol{\lambda}_{\mathbf{p}}[p]$ for $0 \leq p \leq P-1$, and $\boldsymbol{\lambda}[d] = 0$ if $(D/P) \nmid d$. Notice that the objective becomes:

$$\begin{aligned}
2\sum_{d=0}^{D} |\boldsymbol{\lambda}[d]| \cdot |\widehat{\mathbf{w}}[d]| &= 2\sum_{p=0}^{P-1} |\boldsymbol{\lambda}[(D/P) \cdot p]| \cdot |\widehat{\mathbf{w}}[(D/P) \cdot p]| \\
&= 2\frac{D}{P} \sum_{p=0}^{P-1} |\boldsymbol{\lambda}_{\mathbf{p}}[p]| \cdot |\widehat{\mathbf{p}}[p]| \\
&= 2\frac{D}{P} \sum_{p=0}^{P-1} |\boldsymbol{\lambda}_{\mathbf{p}}[p]| \cdot |\widehat{\mathbf{p}}[p]| \\
&= \frac{D}{P} \mathcal{R}^{\mathrm{SDP}}_K(\mathbf{p}) \\
&= \frac{D}{P} \mathcal{R}_{K,C}(\mathbf{p}),
\end{aligned}$$

where the last equality follows from tightness of the SDP (Theorem 4). It thus suffices to show that $\boldsymbol{\lambda}$ is dual feasible. For $\mathbf{x} \in \mathbb{C}^K$ such that $\|\mathbf{x}\| \leq 1$, consider:

$$\begin{aligned}
\sum_{d=0}^{D-1} |\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 &= \sum_{p=0}^{P-1} |\widehat{\mathbf{x}}[(D/P) \cdot p]|^2 \cdot |\boldsymbol{\lambda}[(D/P) \cdot p]|^2 \\
&= \frac{D}{P} \sum_{p=0}^{P-1} |\widehat{\mathbf{x}}[(D/P) \cdot p]|^2 \cdot |\boldsymbol{\lambda}_{\mathbf{p}}[p]|^2.
\end{aligned}$$

Now, let $\widehat{x}^{(P)} \in \mathbb{C}^P$ be the Fourier representation of x when the base dimension is P. Observe that $\widehat{\mathbf{x}}[(D/P) \cdot p]$ is equal to $\sqrt{\frac{P}{D}}\widehat{\mathbf{x}}^{(P)}[p]$. Thus the above expression is equal to:

$$\frac{D}{P}\frac{P}{D}\sum_{p=0}^{P-1} |\widehat{\mathbf{x}}^{(P)}[p]|^2 \cdot |\boldsymbol{\lambda}_{\mathbf{p}}[p]|^2 = \sum_{p=0}^{P-1} |\widehat{\mathbf{x}}^{(P)}[p]|^2 \cdot |\boldsymbol{\lambda}_{\mathbf{p}}[p]|^2.$$

Since $\boldsymbol{\lambda}_{\mathbf{p}}[p]$ is dual feasible for the P-dimensional problem, we see that this is at most 1, as desired.

Case 2: $\mathbf{K} = \mathbf{T} \cdot \mathbf{P}$. We consider $\boldsymbol{\lambda}[(D/P) \cdot p] = \frac{\sqrt{D}}{\sqrt{K}}$ for $0 \le p \le P-1$, and $\boldsymbol{\lambda}[d] = 0$ if $(D/P) \nmid d$. Notice that the objective in eq. (23) is equal to:

$$2\sum_{d=0}^{D} |\boldsymbol{\lambda}[d]| \cdot |\widehat{\mathbf{w}}[d]| = 2\frac{\sqrt{D}}{\sqrt{K}}\sum_{p=0}^{P-1} |\widehat{\mathbf{w}}[(D/P) \cdot p]| = 2\frac{\sqrt{D}}{\sqrt{K}}\|\widehat{\mathbf{w}}\|_1,$$

as desired. It thus suffices to show that that $\boldsymbol{\lambda}$ is dual feasible. For $\mathbf{x} \in \mathbb{C}^K$ such that $\|\mathbf{x}\| \le 1$, we consider

$$\sum_{d=0}^{D-1} |\widehat{\mathbf{x}}[d]|^2 \cdot |\boldsymbol{\lambda}[d]|^2 = \frac{D}{K}\sum_{p=0}^{P-1} |\widehat{\mathbf{x}}[(D/P) \cdot p]|^2.$$

Let $\mathbf{x}_0, \ldots, \mathbf{x}_{T-1} \in \mathbb{C}^P$ be defined so that $\mathbf{x}_t[p] = \mathbf{x}[t \cdot K + p]$ for $0 \le p \le P-1$ and $0 \le t \le T-1$. Now, let $\widehat{\mathbf{x}}_t^{(P)}$ denote the Fourier representation of $\mathbf{x}$ when the base dimension is P, and observe that $\widehat{\mathbf{x}}[(D/P) \cdot p] = \sqrt{\frac{P}{D}}\sum_{t=0}^{T-1} \widehat{\mathbf{x}}_t^{(P)}[p]$. Thus we can rewrite the above expression as:

$$\begin{aligned}
\frac{D}{K}\frac{P}{D}\sum_{p=0}^{P-1} \left|\sum_{t=0}^{T-1} \widehat{\mathbf{x}}_t^{(P)}[p]\right|^2 &= \frac{1}{T}\sum_{p=0}^{P-1}\sum_{t=0}^{T-1} T\left|\widehat{\mathbf{x}}_t^{(P)}[p]\right|^2 \\
&= \sum_{t=0}^{T-1} \left\|\widehat{\mathbf{x}}_t^{(P)}\right\|^2 \\
&= \sum_{t=0}^{T-1} \|\mathbf{x}_t\|^2 \\
&= \|\mathbf{x}\|^2 \le 1,
\end{aligned}$$

as desired. This completes the proof. $\square$

From these matching upper and lower bounds, we can easily conclude Lemma 8.

Proof of Lemma 8. Lemma 8 follows directly from Lemma 17 and Lemma 18. $\square$

E Appendix for Section 5: Networks with multi-channel inputs

For all the results in this appendix, we recall that the weights of the first and second layer are denoted as $\boldsymbol{\mathcal{U}} = \{\mathbf{U}_r\}_{r\in[R]}$ with $\mathbf{U}_r \in \mathbb{R}^{K\times C}\forall_{r\in[R]}$ and $\mathbf{V} \in \mathbb{R}^{D\times C}$, respectively.

E.1 Realizability of linear functions

We show that multiple output channels can be needed to merely realize all linear maps for multi-input-channel networks.

Lemma 9. *For any K, C and R, in order for the the network represented by $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ in eq.* (12) *to realizes all linear maps in $\mathbb{R}^{D\times R}$ it is necessary that $K \cdot C \geq \min\{R, D\}$.*

We prove this lemma by showing that the sub-network corresponding to each output channel can realize a matrix in $\mathbb{R}^{D\times R}$ of rank at most K.

Proof. We first reiterate the expressions for the linear predictor $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ in terms of $\boldsymbol{\mathcal{U}}, \mathbf{V}$:

$$\forall_{r\in[R]},\ W(\boldsymbol{\mathcal{U}}, \mathbf{V})[:, r] = \sum_{c=0}^{C-1} \Big(\mathbf{U}_r[:, c] \star \mathbf{V}[:, c]^{\downarrow}\Big)^{\downarrow}. \tag{24}$$

We will now express the above formulation as matrix multiplication using the following new notation: $\forall_{c\in[C]}$ let $\underline{\mathbf{U}}_c \in \mathbb{R}^{K\times R}$ denote the representation of first layer weights corresponding to each output channel such that

$$forall_{r\in[R]} \forall_{c\in[C]},\ \underline{\mathbf{U}}_c[:, r] = \mathbf{U}_r[:, c].$$

For $c \in [C]$, consider the following matrix which consists of first K columns of the circulant matrix formed by $\mathbf{V}[:, c]$:

$$\forall_{c\in[C]},\ \widetilde{\mathbf{V}}_c = \frac{1}{\sqrt{D}} \begin{bmatrix} \mathbf{V}[0, c] & \mathbf{V}[D-1, c] & \cdots & \mathbf{V}[D-K+1, c] \\ \mathbf{V}[1, c] & \mathbf{V}[0, c] & \cdots & \mathbf{V}[D-K+2, c] \\ \vdots & \vdots & & \vdots \\ \mathbf{V}[D-1, c] & \mathbf{V}[D-2, c] & \cdots & \mathbf{V}[D-K, c] \end{bmatrix} \in \mathbb{R}^{D\times K}$$

Based on this notation, we can check by following the definitions that for all c,

$$\Big(\mathbf{U}_r[:, c] \star \mathbf{V}[:, c]^{\downarrow}\Big)^{\downarrow} = \widetilde{\mathbf{V}}_c \mathbf{U}_r[:, c] = \widetilde{\mathbf{V}}_c\, \underline{\mathbf{U}}_c[:, r].$$

We can thus write $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ as follows:

$$W(\boldsymbol{\mathcal{U}}, \mathbf{V}) = \sum_{c=0}^{C-1} \widetilde{V}_c\, \underline{\mathbf{U}}_c. \tag{25}$$

We now observe that each term in the summation $\widetilde{V}_c\, \underline{\mathbf{U}}_c$ is of rank utmost K as $\widetilde{V}_c \in \mathbb{R}^{D\times K}$ and $\underline{\mathbf{U}}_c \in \mathbb{R}^{D\times K}$. Thus, for any $\boldsymbol{\mathcal{U}}, \mathbf{V}$, $\mathrm{rank}(W(\boldsymbol{\mathcal{U}}, \mathbf{V})) \leq K \cdot C$. From this we conclude that in order to realize all linear maps in the multi-channel input space of $\mathbb{R}^{D\times R}$, we necessarily need $K \cdot C \geq \min\{R, D\}$.

Additionally, in eq. (25) we see that since $\underline{\mathbf{U}}_c$ are unconstrained, each term in the sum can realize any rank 1 matrix. This implies that $C \geq \min\{R, D\}$ is a sufficient condition for $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$ to realize any $\mathbf{W} \in \mathbb{R}^{D\times R}$. However, from Theorem 13 we know that this condition is not necessary. It is an open question to derive the tightest necessary and sufficient conditions.

We finally note that a similar proof can be shown using the Fourier representation in eq. (12). □

However, in the special cases of $K = 1$ and $K = D$, in Theorems 12-13 we show that the SDP is tight once C is large enough to realize all linear functions, which in these cases is $C \geq R/K$. In these end cases, we further derive interesting closed form expressions of $\mathcal{R}_{K,C,R}(\mathbf{W})$.

E.2 Proof of Lemma 10

Lemma 10. *For any* $\mathbf{W} \in \mathbb{R}^{D\times R}$*, and any* $C \geq RK$*, it holds that* $\mathcal{R}_{K,C,R}(\mathbf{W}) = \mathcal{R}^{SDP}_{K,R}(\mathbf{W})$*.*

Proof. We show a stronger statement: *any* optimal solution to the SDP has rank at most RK. This implies the desired result, because $\mathcal{R}_{K,C,R}(\mathbf{W})$ is equivalent to the SDP with a rank constraint of C.

For the remainder of the proof, we let $\mathbf{Z}$ be an optimal solution to the SDP and we prove that $\text{rank}(\mathbf{Z}) \leq RK$. Let $\text{rank}(\mathbf{Z}) = L$. We can write:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{U}_0 \\ \mathbf{U}_1 \\ \dots \\ \mathbf{U}_R \\ \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{U}_0^\top & \mathbf{U}_1^\top & \dots & \mathbf{U}_R^\top & \mathbf{V}^\top \end{bmatrix}$$

where the matrices $\mathbf{U}_r \in \mathbb{R}^{K\times L}$ for $0 \leq r \leq R-1$ correspond to the weights in the convolution layer, the matrix $\mathbf{V} \in \mathbb{R}^{D\times L}$ corresponds to the weights in the linear layer. It suffices to show that there exists a spanning set of the column space of

$$\begin{bmatrix} \mathbf{U}_0 \\ \mathbf{U}_1 \\ \dots \\ \mathbf{U}_R \\ \mathbf{V} \end{bmatrix}$$

that has at most RK elements.

The key ingredient of the proof is the KKT conditions. Using similar logic to Appendix C, we can write the KKT conditions in the following form:

$$\text{for all } 0 \leq r \leq R-1 : \widehat{\mathbf{W}}[:, r] = \widehat{\mathbf{U}}_r \odot \widehat{\mathbf{V}} \tag{KKT 1}$$

$$\left[\begin{array}{cccc:c} & & & & \overline{\mathbf{F}}_K^T \mathbf{\Lambda}_0 \overline{\mathbf{F}} \\ & \mathbf{0}_{(R\cdot K)\times(R\cdot K)} & & & \overline{\mathbf{F}}_K^T \mathbf{\Lambda}_1 \overline{\mathbf{F}} \\ & & & & \vdots \\ & & & & \overline{\mathbf{F}}_K^T \mathbf{\Lambda}_{R-1} \overline{\mathbf{F}} \\ \hdashline \mathbf{F}\overline{\mathbf{\Lambda}}_0\mathbf{F}_K & \mathbf{F}\overline{\mathbf{\Lambda}}_1\mathbf{F}_K & \dots & \mathbf{F}\overline{\mathbf{\Lambda}}_{R-1}\mathbf{F}_K & \mathbf{0}_{D\times D} \end{array}\right] \preceq \mathbf{I}_{D+KR} \tag{KKT 2}$$

$$\overline{\widehat{\mathbf{V}}} = \sum_{r=0}^{R-1} \overline{\mathbf{\Lambda}}_r \widehat{\mathbf{U}}_r \tag{KKT 3}$$

$$\text{for all } 0 \leq r \leq R-1 : \widehat{\mathbf{U}}_r = \mathbf{F}_K \overline{\mathbf{F}}_K^T \mathbf{\Lambda}_r \overline{\widehat{\mathbf{V}}}. \tag{KKT 4}$$

where the matrices $\mathbf{\Lambda}_r \in \mathbb{C}^{D\times D}$ for $0 \leq r \leq R-1$ are diagonal and correspond to the relevant dual variables. Let's use (KKT 4) to construct a spanning set of the column space of this matrix. Let $S_r \left\{\mathbf{u}_r \in \mathbb{R}^K\right\}$ be a basis for the column space of $\mathbf{U}_r$ for each $0 \leq r \leq R$; since $\mathbf{U}_r$ is $K \times L$ dimensional, we see that S_r has at most K elements. We see by (KKT 4) that the set of vectors given by the concatenation of $[\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{R-1}, \overline{F}\left(\sum_{r=0}^{R-1} \overline{\mathbf{\Lambda}}_r \widehat{\mathbf{u}}_r]\right)]$ for $\mathbf{u}_r \in S_r$ for each $0 \leq r \leq R$ spans the column space of the desired matrix. By construction, this spanning set has at most RK elements, as desired.

□

E.3 Proofs of Theorems 12-13: induced regularizer for $K = 1$ and $K = D$

Theorem 12. *For any $\mathbf{W} \in \mathbb{R}^{D\times R}$, and any $C \geq \min\{R, D\}$, the induced regularizer for $K = 1$ is given by the scaled nuclear norm $\|.\|_*$:*

$$\mathcal{R}_{1,C,R}(\mathbf{W}) = 2\sqrt{D}\|\mathbf{W}\|_* = 2\sqrt{D}\|\widehat{\mathbf{W}}\|_*.$$

Proof. For $K = 1$, we have that $\forall_{r\in[R]}$, $\mathbf{U}_r \in \mathbb{R}^{1\times C}$. For this proof, we stack the vectors $\mathbf{U}_r$ to obtain $\widetilde{\mathbf{U}} \in \mathbb{R}^{R\times C}$ such that $\forall_{r\in[R]}$, $\widetilde{\mathbf{U}}[r, :] = \mathbf{U}_r$.

We work with the signal space definition of the linear predictor realized by network as: $\forall_{r\in[R]}$, $W(\boldsymbol{\mathcal{U}}, \mathbf{V})[:, r] = \sum_{c=0}^{C-1} \big(\mathbf{U}_r[:, c] \star \mathbf{V}[:, c]^{\downarrow}\big)^{\downarrow}$ (from eq. (12)).

For kernel size of 1, we notice that from the definition of convolution in Definition 1, we have the following:

$$\Big(\mathbf{U}_r[:, c] \star \mathbf{V}[:, c]^{\downarrow}\Big)^{\downarrow} = \frac{\widetilde{\mathbf{U}}[r, c]}{\sqrt{D}}\mathbf{V}[:, c] \propto \mathbf{V}[:, c]. \tag{26}$$

Plugging this back into the expression of $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$, we have the following:

$$W(\boldsymbol{\mathcal{U}}, \mathbf{V})[:, r] = \frac{1}{\sqrt{D}}\mathbf{V}\widetilde{\mathbf{U}}[r, :] \implies W(\boldsymbol{\mathcal{U}}, \mathbf{V}) = \frac{1}{\sqrt{D}}\mathbf{V}\widetilde{\mathbf{U}}^\top. \tag{27}$$

In the above formulation $\mathbf{V} \in \mathbb{R}^{D\times C}, \widetilde{\mathbf{U}} \in \mathbb{R}^{R\times C}$ are of rank C, but otherwise completely unconstrained. Thus, they can realize any rank K matrix. So as long as $C \geq \min\{R, D\}$, the network can realize any linear predictor $\mathbf{W} \in \mathbb{R}^{D\times R}$.

The rest of the proof follows from connecting the above expression into the variational characterization of the nuclear norm. The induced regularizer $\mathcal{R}_{1,C,R}(\mathbf{W})$ from eq. (13) for $C \geq \min\{R, D\}$ can now be expressed as follows:

$$\begin{aligned} \mathcal{R}_{1,C,R}(\mathbf{W}) = \min_{\widetilde{\mathbf{U}}\in\mathbb{R}^{R\times C}, \mathbf{V}\in\mathbb{R}^{D\times C}} & \quad \|\widetilde{\mathbf{U}}\|^2 + \|\mathbf{V}\|^2 \\ \text{s.t.,} & \quad \sqrt{D}\mathbf{W} = \mathbf{V}\widetilde{\mathbf{U}}^\top. \end{aligned} \tag{28}$$

For $C \geq \min\{R, D\}$ eq. (28) is exactly the variational definition of nuclear norm (see Rennie and Srebro [Lemma 1 RS05]) and thus $\mathcal{R}_{1,C,R}(\mathbf{W}) = 2\sqrt{D}\|\mathbf{W}\|_*$ for even unbounded C. The fact that $C = \min\{R, D\}$ is sufficient can be seen by obtaining the optimum nuclear norm as upper bound from using $\mathbf{V} = \mathbf{L}\sqrt{\boldsymbol{\Sigma}}$ and $\widetilde{\mathbf{U}} = \mathbf{R}\sqrt{\boldsymbol{\Sigma}}$, where $\sqrt{D}\mathbf{W} = \mathbf{L}\boldsymbol{\Sigma}\mathbf{R}^\top$ is the singular value decomposition of $\sqrt{D}\mathbf{W}$. Finally, we note that based on our normalization of Fourier transform, we have $\|\mathbf{W}\|_* = \|\widehat{\mathbf{W}}\|_*$. This completes our proof.

Note: For $K = 1$ we provided the proof of $\mathcal{R}_{1,C,R}(\mathbf{W})$ in the signal space of $\mathbf{W}$, but the Theorem can also be proved in the Fourier domain (similar to the proof of Theorem 13 given below) by first showing that the SDP relaxation evaluates to the nuclear norm and combining this with the matching upper bound for $\mathcal{R}_{1,C,R}(\mathbf{W})$ shown above. □

Theorem 13. *For any $\mathbf{W} \in \mathbb{R}^{D\times R}$, and any $C \geq 1$, the induced regularizer for $K = D$ is given as follows*

$$\mathcal{R}_{D,C,R}(\mathbf{W}) = 2\|\widehat{\mathbf{W}}\|_{2,1} := \sum_{d=0}^{D-1}\sqrt{\sum_{r=0}^{R-1}\big|\widehat{\mathbf{W}}[d, r]\big|^2}.$$

Proof. We begin by expressing induced regularizer for full dimensional kernel sizes $\mathcal{R}_{D,C,R}(\mathbf{W})$ from eq. (13) in terms of the Fourier representation of the linear predictor realized by the network $W(\boldsymbol{\mathcal{U}}, \mathbf{V})$:

$$\begin{aligned} \mathcal{R}_{D,C,R}(\mathbf{W}) = \inf_{\boldsymbol{\mathcal{U}},\mathbf{V}} \quad & \|\widehat{\mathbf{U}}_r\|^2 + \|\widehat{\mathbf{V}}\|^2 \\ \text{s.t.,} \quad & \forall_{r\in[R]} \mathbf{W}[:, r] = \operatorname{diag}(\widehat{\mathbf{U}}_r \widehat{\mathbf{V}}^\top), \end{aligned} \tag{29}$$

where $\boldsymbol{\mathcal{U}}, \mathbf{V}$ are of dimensions $\boldsymbol{\mathcal{U}} = \{\mathbf{U}_r \in \mathbb{R}^{D\times C}\}_{r\in[R]}, \mathbf{V} \in \mathbb{R}^{D\times C}$.

Our proof for the case of networks with multi-channel inputs with $K = D$ follows the following structure:

Step 1. We first show an upper bound on the induced regularizer for single output channel $C = 1$ with full dimensional kernel $K = D$ as $\mathcal{R}_{D,1,R}(\mathbf{W}) \leq 2\|\widehat{\mathbf{W}}\|_{2,1}$ by providing a construction of $\boldsymbol{\mathcal{U}}, \mathbf{V}$. It immediately follows from the monotonicity of $\mathcal{R}_{D,C,R}$ that for all C, $\mathcal{R}_{D,C,R}(\mathbf{W}) \leq \mathcal{R}_{D,1,R}(\mathbf{W}) \leq 2\|\widehat{\mathbf{W}}\|_{2,1}$. This step also consequently shows that when $K = D$, every linear predictor over the multi-channel input space of $\mathbb{R}^{D\times R}$ is realizable by a network with even a single output channel.

Step 2. The bulk of our proof lies in matching the upper bound with a lower bound on the dual problem of the SDP in eq. (14) as $\mathcal{R}^{\text{SDP}}_{D,R}(\mathbf{W}) \geq 2\|\widehat{\mathbf{W}}\|_{2,1}$. This gives us that for all $C \geq 1$, $\mathcal{R}_{D,C,R}(\mathbf{W}) \geq \mathcal{R}^{\text{SDP}}_{D,R}(\mathbf{W}) \geq 2\|\widehat{\mathbf{W}}\|_{2,1}$.

Step 1. Upper bound on the induced regularizer $\boldsymbol{\mathcal{R}_{D,C,R}}$ We first show that $\mathcal{R}_{D,1,R}(\mathbf{W}) \leq 2\|\widehat{\mathbf{W}}\|_{2,1} = 2\sum_{d\in[D]} \|\widehat{\mathbf{W}}[d, :]\|$. Since $\mathcal{R}_{D,R,C}(\mathbf{W})$ is decreasing in C, it suffices to show this for $C = 1$. For $C = 1$ and $K = D$, we have $\mathbf{V} \in \mathbb{R}^D$ and $\forall_{r\in[R]}\ \mathbf{U}_r \in \mathbb{R}^D$. For full dimensional kernels, the Fourier domain representations $\widehat{\mathbf{U}}_0, \widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_R, \widehat{\mathbf{V}} \in \mathbb{C}^D$ are all unconstrained beyond the symmetry properties of Fourier transform of real matrices. Thus consider the following $\boldsymbol{\mathcal{U}}, \mathbf{V}$:

$$\forall_{d\in[D]}\forall_{r\in[R]}, \widehat{\mathbf{U}}_r[d] = \frac{\widehat{\mathbf{W}}[d, r]}{\sqrt{\|\widehat{\mathbf{W}}[d, :]\|}}, \quad \text{and} \quad \widehat{\mathbf{V}}[d] = \sqrt{\|\widehat{\mathbf{W}}[d, :]\|}. \tag{30}$$

It is easy to see that the above $\boldsymbol{\mathcal{U}}, \mathbf{V}$ satisfy the constraints of $\mathcal{R}_{D,1,R}(\mathbf{W})$ in eq. (29) that is $\forall_{r\in[R]}\mathbf{W}[:, r] = \operatorname{diag}(\widehat{\mathbf{U}}_r \widehat{\mathbf{V}}^\top)$. Further $\widehat{\mathbf{U}}_0, \widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_R, \widehat{\mathbf{V}} \in \mathbb{C}^D$ satisfy the required symmetry properties since $\mathbf{W}$ is real.

Now computing the objective, we immediately have that $\|\widehat{\mathbf{V}}\|^2 = \sum_{d\in[D]} \|\widehat{\mathbf{W}}[d, :]\|$, and further,

$$\sum_{r\in[R]} \|\widehat{\mathbf{U}}_r\|^2 = \sum_{d\in[D]} \left[\sum_{r\in[R]} \frac{|\widehat{\mathbf{W}}[d, r]|^2}{\|\widehat{\mathbf{W}}[d, :]\|}\right] = \sum_{d\in[D]} \|\widehat{\mathbf{W}}[d, :]\|.$$

This construction thus gives us the desired upper bound

$$\mathcal{R}_{D,C,R}(\mathbf{W}) \leq \mathcal{R}_{D,1,R}(\mathbf{W}) \leq 2\|\widehat{\mathbf{W}}\|_{2,1}. \tag{31}$$

Step 2: Lower bound on the induced regularizer $\boldsymbol{\mathcal{R}_{D,C,R}}$ We show the lower bound by lower bounding the dual problem of the SDP in eq. (14).

For $r \in [R]$, let $\boldsymbol{\lambda}^{\text{real}}_r \in \mathbb{R}^D$ and $\boldsymbol{\lambda}^{\text{img}}_r \in \mathbb{R}^D$ denote the dual variables corresponding to the constraints in the SDP in eq. (14) for the real and imaginary parts, respectively, of $\widehat{\mathbf{W}}[:, r]$. Similar to single input channel proofs, we define $\boldsymbol{\lambda}_r = \boldsymbol{\lambda}^{\text{real}}_r + i \cdot \boldsymbol{\lambda}^{\text{img}}_r$ and $\boldsymbol{\Lambda}_r = \operatorname{diag}(\boldsymbol{\lambda}_r)$. Additionally, we introduce the notation for the matrix obtained by taking $\{\boldsymbol{\lambda}_r\}_r$ as columns: $\Xi = [\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_R] \in \mathbb{C}^{D\times R}$ such that $\forall_r$, $\Xi[:, r] = \boldsymbol{\lambda}_r$.

Based on weak duality for the SDP in eq. (14), we have the following:

$$\mathcal{R}^{\text{SDP}}_{D,R}(\mathbf{W}) \geq \begin{array}{cl} \max\limits_{\Xi=[\boldsymbol{\lambda}_0,\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_R]} & 2\cdot \mathrm{Re}(\langle \widehat{\mathbf{W}}, \Xi\rangle) \\ \text{s.t.,} & \sum\limits_{d,r}(\boldsymbol{\lambda}_r^{\text{real}}[d]\cdot \mathbf{A}^{\text{real}}_{d,r} + \boldsymbol{\lambda}_r^{\text{img}}[d]\cdot \mathbf{A}^{\text{img}}_{d,r}) \preccurlyeq \mathbf{I}. \end{array} \tag{32}$$

Our rest of the proof obtains the lower bound by constructing an appropriate Ξ satisfying the constraints:

From the definitions of $\{\mathbf{A}^{\text{real}}_{d,r}, \mathbf{A}^{\text{img}}_{d,r}\}_{d\in[D],r\in R}$ in Appendix 5.2 and using $\mathbf{Q}_d = \overline{\mathbf{F}}\mathrm{e}_d\mathrm{e}_d^\top\overline{\mathbf{F}}$ (note that for $K = D$ and $\mathbf{F}_K = \mathbf{F} = \mathbf{F}^\top$), we have the following:

$$\sum_{d,r}(\boldsymbol{\lambda}_r^{\text{real}}[d]\cdot \mathbf{A}^{\text{real}}_{d,r} + \boldsymbol{\lambda}_r^{\text{img}}[d]\cdot \mathbf{A}^{\text{img}}_{d,r}) = \left[\begin{array}{cccc:c} & & & & \overline{\mathbf{F}}\boldsymbol{\Lambda}_0\overline{\mathbf{F}} \\ & \mathbf{0}_{(R\cdot D)\times(R\cdot D)} & & & \overline{\mathbf{F}}\boldsymbol{\Lambda}_1\overline{\mathbf{F}} \\ & & & & \vdots \\ & & & & \overline{\mathbf{F}}\boldsymbol{\Lambda}_R\overline{\mathbf{F}} \\ \hdashline \mathbf{F}\overline{\boldsymbol{\Lambda}}_0\mathbf{F}_K & \mathbf{F}\overline{\boldsymbol{\Lambda}}_1\mathbf{F}_K & \ldots & \mathbf{F}\overline{\boldsymbol{\Lambda}}_R\mathbf{F}_K & \mathbf{0}_{D\times D}\end{array}\right] \tag{33}$$

We state and prove the following claim:

Claim. Ξ *satisfies the constraints of the dual problem in the RHS of eq.* (32) *if* $\max\limits_{d\in[D]} \|\Xi[d,:]\| \leq 1$

Proof of claim. The relevant constraint in eq. (32) is $\sum\limits_{d,r}(\boldsymbol{\lambda}_r^{\text{real}}[d]\cdot \mathbf{A}^{\text{real}}_{d,r} + \boldsymbol{\lambda}_r^{\text{img}}[d]\cdot \mathbf{A}^{\text{img}}_{d,r}) \preccurlyeq \mathbf{I}$. It thus suffices to show that for all $\mathbf{y} \in \mathbb{R}^D$ such that $\|\mathbf{y}\| = 1$, it holds that $\sum_{r=0}^R \|\overline{\mathbf{F}}\boldsymbol{\Lambda}_r\overline{\mathbf{F}}\mathbf{y}\|^2 \leq 1$. We have the following set of inequalities that prove the claim $\forall_{\mathbf{y}:\|\mathbf{y}\|=1}$:

$$\begin{aligned}\sum_{r=0}^R \|\overline{\mathbf{F}}\boldsymbol{\Lambda}_r\overline{\mathbf{F}}\mathbf{y}\|^2 &\overset{(a)}{=} \sum_{r=0}^R \|\boldsymbol{\Lambda}_r(\overline{\mathbf{F}}\mathbf{y})\|^2 \\ &= \sum_{d=0}^D\sum_{r=0}^R |\boldsymbol{\lambda}_r[d]|^2\cdot|(\overline{\mathbf{F}}\mathbf{y})[d]|^2 \overset{(b)}{=} \sum_{d=0}^D \|\Xi[d,:]\|^2\cdot|(\overline{\mathbf{F}}\mathbf{y})[d]|^2 \\ &\leq \max_{d\in[D]}\|\Xi[d,:]\|^2\|\overline{\mathbf{F}}\mathbf{y}\|^2 \\ &\overset{(c)}{=} \max_{d\in[D]}\|\Xi[d,:]\|^2,\end{aligned}$$

where (a) follows from $\overline{\mathbf{F}}$ being unitary, (b) from definition of Ξ, and (c) from $\|\mathbf{F}\mathbf{y}\| = \|\mathbf{y}\| = 1$. □

Consider Ξ defined as follows:

$$\forall_{d\in[D]}, \quad \Xi[d,:] = \frac{\widehat{\mathbf{W}}[d,:]}{\|\widehat{\mathbf{W}}[d,:]\|}$$

It is easy to check that $\max\limits_{d\in[D]}\|\Xi[d,:]\| = 1$ and thus based on the claim we proved above Ξ satisfies the constraints of the dual optimization problem in the RHS of eq. (32). Additionally, the objective evaluates to the desired bound of $\mathrm{Re}(\langle\widehat{\mathbf{W}}, \Xi\rangle) = \sum_d \|\widehat{\mathbf{W}}[d,:]\| = \|\widehat{\mathbf{W}}\|_{2,1}$. We thus have the following lower bound:

$$\mathcal{R}_{D,C,R}(\mathbf{W}) \geq \mathcal{R}^{\text{SDP}}_{D,R}(\mathbf{W}) \geq 2\,\mathrm{Re}(\langle\widehat{\mathbf{W}}, \Xi\rangle) = 2\|\widehat{\mathbf{W}}\|_{2,1}. \tag{34}$$

Conclusion of the proof. The proof of the Theorem follows from combining the matching upper and lower bounds in eq. (31) and eq. (34), respectively, to obtain $\mathcal{R}_{D,C,R}(\mathbf{W}) = 2\|\widehat{\mathbf{W}}\|_{2,1}$. $\square$