Geometry of Linear Convolutional Networks[*]

Kathlén Kohn[†], Thomas Merkh[‡], Guido Montúfar[§], and Matthew Trager[¶]

Abstract. We study the family of functions that are represented by a linear convolutional network (LCN). These functions form a semi-algebraic subset of the set of linear maps from input space to output space. In contrast, the families of functions represented by fully connected linear networks form algebraic sets. We observe that the functions represented by LCNs can be identified with polynomials that admit certain factorizations, and we use this perspective to describe the impact of the network's architecture on the geometry of the resulting function space. We further study the optimization of an objective function over an LCN, analyzing critical points in function space and in parameter space and describing dynamical invariants for gradient descent. Overall, our theory predicts that the optimized parameters of an LCN will often correspond to *repeated filters* across layers, or filters that can be decomposed as repeated filters. We also conduct numerical and symbolic experiments that illustrate our results and present an in-depth analysis of the landscape for small architectures.

1. Introduction. A neural network is a parameterized family of functions. The class of representable functions, also known as the *neuromanifold* or *function space*, is determined by the architecture of the network. Optimizing an objective function over such a parametrized set usually depends on both the set and the parameterization. The objective function is typically convex in function space but nonconvex in parameter space. This observation has guided numerous recent works on parameter optimization in neural networks; see, e.g., [23, 16, 1, 17]. However, the geometry of function space is not well understood in general. Fully connected networks, also known as dense networks, have been studied in significantly more detail than convolutional networks, which are the topic of this article. Whereas weight matrices in fully connected networks have independent and unconstrained entries, weight matrices in convolu-

[†]Department of Mathematics, KTH, Stockholm, 10044, Sweden (kathlen@kth.se).
[‡]Department of Mathematics, UCLA, Los Angeles, CA 90095 USA (tmerkh@g.ucla.edu).
[§]Departments of Mathematics and Statistics, UCLA, Los Angeles, CA 90095 USA; Max Planck Institute for Mathematics in the Sciences, Leipzig, Saxony, 04103, Germany (montufar@math.ucla.edu).
[¶]Amazon, New York, NY USA 10001 (matthewtrager@gmail.com). This work was done outside of Amazon.

tional networks are sparse and have repeated entries.

Neural networks with linear activation functions, called linear networks, represent linear functions as compositions of linear functions. An important motivation for studying such networks is to better understand the effects of the parameterization when optimizing an objective function. We will see that it also can help us better understand the effects that the geometry of function space have on the optimization problem. Although neural networks are usually presented in a parameterized form, it is in principle possible to characterize the representable functions implicitly as the solutions to certain constraints. Such descriptions have been particularly fruitful in algebraic statistics; see, e.g., [34, 32, 2, 12, 13]. In the case of fully connected linear networks, the function space is characterized by rank constraints depending on the layer widths. A study of critical points and local optimizers in parameter space and function space for fully connected linear networks has been carried out in [35]. Here we follow a similar perspective, analyzing the function space and the optimization of parameters for *linear convolutional networks* (LCNs).

Main results. In section 2, we introduce LCNs and present our general setup. Each layer of an LCN is represented by a convolutional matrix or, more generally, a convolutional tensor. We observe that composing convolutional layers corresponds to polynomial multiplication.

In section 3, we recall some general properties of discriminants and multiplicity root patterns for polynomials, which we use in later sections.

In section 4, we show that the function space of an LCN (i.e., the set of linear maps that can be represented by the network) is a semi-algebraic set, whose natural ambient space is the vector space of convolutional tensors of a certain format. In contrast to fully connected linear networks, whose function space is cut out by polynomial equations only, the function space of LCNs is determined by polynomial equations and inequalities. For one-dimensional convolutions with stride one, the function space is a full-dimensional subset of its natural ambient space and its boundary is contained in the discriminant hypersurface of univariate polynomials of a certain degree. The function space is equal to its natural ambient space if and only if at most one of the filters has even size. The latter case is interesting since optimizing a convex loss in function space reduces to a convex optimization problem on a vector space. For higher-dimensional convolutions or larger strides, the function space is typically a lower-dimensional subset of its natural ambient space.

In section 5, we study the optimization problem for a general loss and one-dimensional convolutions with stride one. Different LCNs can have the same function space, but different parameterizations can influence the optimization behavior. Hence, we distinguish *pure critical points* given by the optimization on function space from *spurious critical points* induced only by the parameterization. We show that the critical points in parameter space are either global minima or correspond to polynomials with repeated roots. The latter actually correspond to critical points of the loss in function space restricted to subsets of polynomials with particular real root multiplicity patterns that we characterize combinatorially depending on the LCN architecture. We further describe the gradient flow dynamics in parameter space and provide invariants that allow us to formulate it as a local Riemannian gradient in function space.

In section 6, we take a closer look at the square loss. We provide upper bounds on the number of critical points in function space in terms of *Euclidean distance degrees* and observe that for certain normalizations of the training data the square loss becomes the *Bombieri*

norm, which is expected to have fewer critical points.

In section 7, we conduct numerical experiments illustrating our theoretical results. These demonstrate how the type of training data, the geometry of function space, and the parameterization lead to different types of critical points and solutions found using gradient descent.

Overview of previous works on linear networks. Loss surface. The properties of the optimization problem in linear networks have been studied in [8, 7, 9]. For the square loss in fully connected linear networks, there is a unique local and global minimum, up to equivalence, and all other critical points are saddles. A proof of this statement for the deep case was given in [24]. For deep residual linear networks, [21] showed that the square loss has no critical points other than global minima. Further, [29, 37] argue that depth alone does not create bad local minima, although it induces a nonconvex loss surface. The analytic form of the critical points for the square loss in fully connected linear networks was discussed in [38]. For arbitrary differentiable convex losses and fully connected linear networks having layers at least as wide as the input or output layers, [26] showed that all local minima are global. A geometric analysis was given in [35], showing that the absence of nonglobal local minima in linear networks is expected for arbitrary smooth convex losses only if the architecture can express all linear maps (as in [26]) or for the quadratic loss and arbitrary architectures (as in [24]). Taking an algebraic standpoint, [30] studied fully connected linear networks and upper bounds on the number of critical points of the square loss with weight norm regularization. The above works focus mostly on fully connected networks. A paper dating back to 1989 [7] asked whether nonglobal local minima for the square loss exist in the case of nonfully connected (i.e., locally connected) multilayer linear networks, and this seems to have remained an open question. We show that LCNs can indeed have nontrivial local minima in function space and in parameter space.

Gradient dynamics. An overview of nonconvex optimization in low-rank matrix factorization (two-layer linear networks) was given in [14]. Exact solutions to the nonlinear gradient dynamics for the square loss in fully connected deep linear networks with orthogonal input data have been obtained in [31]. Also for the square loss in fully connected linear networks, [4] showed linear convergence to a global minimum when the hidden layers are at least as wide as the input or output layers, weight matrices at initialization are approximately balanced, and the initial loss is less than for any rank deficient solution. The work [5] shows that depth amounts to a preconditioner which may accelerate convergence. A detailed analysis of gradient flow optimization and convergence for the square loss in fully connected linear networks was presented in [6]. Using a balancedness condition, they obtain a Riemannian gradient in function space and show that the flow always converges to a critical point, which for almost all initializations is the global optimum in function space restricted to the set of rank-k matrices for some k. We characterize gradient flow invariants in LCNs and obtain corresponding descriptions in function space. The critical points can also be characterized in function space, in terms of real root multiplicity patterns rather than rank constraints.

Linear convolutional networks and implicit bias of gradient descent. LCNs have been studied in [20] for the case of one-dimensional convolutions with circulant weight matrices (with stride 1) having full-width filters and a single output. Such networks can express any linear map. That work focuses on binary classification with linearly separable data and asks which of the infinitely many linear classifiers with zero loss are obtained by gradient descent. For

the exponential loss, it is shown that for fully connected networks gradient descent converges to the hard margin support vector machine linear classifier independently of the depth of the network (as was previously observed for the direct parameterization in [33]), whereas for convolutional networks it is biased towards linear classifiers that are sparse in frequency domain depending on the depth. In contrast, we consider LCNs with arbitrary filter sizes (and also provide some results for LCNs with larger strides or higher-dimensional convolutions). As we will see, the filter sizes play an important role. Our results suggest that gradient descent in LCNs will be biased towards solutions with repeated filters (or repeated factors of filters), even when these are not global optimizers of the training objective.

Notation. Throughout the paper, we use zero-based indices for our vectors and matrices. For $n \in \mathbb{Z}_{\geq 0}$, we write $[n] = \{0, \ldots, n-1\}$. We write $\mathbb{R}[\mathsf{x},\mathsf{y}]_d$ (resp., $\mathbb{C}[\mathsf{x},\mathsf{y}]_d$) for the space of bivariate homogeneous polynomials of degree d with real (resp., complex) coefficients.

2. Linear convolutional networks. A *feedforward neural network* is a family of functions $f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$ that are compositions of linear and nonlinear maps:

$$(2.1) \qquad\qquad f(x) = (\alpha_L \circ \rho \circ \alpha_{L-1} \circ \rho \circ \ldots \rho \circ \alpha_1)(x).$$

Here $\alpha_l : \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_l}$ is an affine map and each $\rho : \mathbb{R}^{d_*} \to \mathbb{R}^{d_*}$ is a (typically nonlinear) *activation map* that acts on each element of a vector. The composition of consecutive affine and activation maps is called a *layer* of the network. The coefficients of the affine functions α_i are called the *weights* and *biases* of the network and serve as trainable parameters.

A *linear network* is a feedforward network where the activation function ρ is the identity. In this setting, the end-to-end map f in (2.1) is affine; however, the parameterization of the representable functions provided by the network's weights is nonlinear. Thus, linear networks preserve some of the nonlinear aspects of general feedforward networks while being theoretically more tractable.

A *linear convolutional network (LCN)* is a linear network with the additional constraint that the affine functions in (2.1) have a convolutional structure. The convolutional structure can be seen as constraining general affine maps to satisfy certain linear equations, corresponding to conditions of *restricted connectivity* (some weight coefficients must be zero) and *weight sharing* (some weight coefficients must be repeated). Although an LCN function f can also be seen as a particular fully connected linear network, the family of all representable functions using an LCN architecture is a strict subset of the family of representable functions using a fully connected architecture. In particular, the study of LCNs cannot be reduced to the study of general linear networks. We will point out several important differences between the geometry and optimization properties of LCNs and those of fully connected networks.

In this paper, we consider convolutional networks with a single filter per layer and no biases. For most of our discussion, we focus on convolutions applied to one-dimensional signals, which means that the input x of the network function is a vector (we discuss the case where x is an image or higher-dimensional in subsection 4.3, but we defer general definitions to that section). In this setting, a *convolution* is a map $\alpha_l : \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_l}$ of the form

$$(2.2) \qquad\qquad (\alpha_l(x))_i = \sum_{j \in [k_l]} w_{l,j} \cdot x_{is_l+j},$$

where $w_l = (w_{l,0}, \ldots, w_{l,k_l-1}) \in \mathbb{R}^{k_l}$ is a *filter* of *width* k_l and $s_l \in \mathbb{Z}_{>0}$ is the *stride* associated with the convolution.[1] We are assuming "no padding" in the way the boundaries of the domain are handled. This tacitly assumes that input dimension and stride fit with the output dimension. We compare this approach with using "circular padding" in subsection 2.3 below.

The linear map α_l in (2.2) can be represented as a generalized Toeplitz matrix (generalized in the sense that standard Toeplitz matrices would correspond to stride $s_l = 1$). For example, for stride $s_l = 2$, filter size $k_l = 3$, and input size $d_{l-1} = 7$, the matrix takes the form

$$(2.3) \qquad W_l = \begin{bmatrix} w_{l,0} & w_{l,1} & w_{l,2} & & & & \\ & & w_{l,0} & w_{l,1} & w_{l,2} & & \\ & & & & w_{l,0} & w_{l,1} & w_{l,2} \end{bmatrix}.$$

We will call generalized Toeplitz matrices *convolutional matrices*. For convolutional matrices as in (2.3), the input and output sizes d_{l-1}, d_l, filter size k_l, and stride s_l are related by

$$(2.4) \qquad d_l = \frac{d_{l-1} - k_l}{s_l} + 1.$$

An L-layer LCN function is given by $f(x) = \overline{W}x$, where $\overline{W} \in \mathbb{R}^{d_L \times d_0}$ can be written as $\overline{W} = W_L W_{L-1} \cdots W_2 W_1$ and each $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ is a convolutional matrix. In this work, we study the *function space*, i.e., the set of linear maps that can be represented by an LCN as the parameters vary:

$$\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} = \left\{ \overline{W} \in \mathbb{R}^{d_L \times d_0} : \overline{W} = \prod_{l=1}^{L} W_l, \ W_l \in \mathbb{R}^{d_l \times d_{l-1}} \text{ convolutional} \right\}.$$

Here $\boldsymbol{d} = (d_0, \ldots, d_L)$ denotes the dimensions, $\boldsymbol{k} = (k_1, \ldots, k_L)$ the filter sizes, and $\boldsymbol{s} = (s_1, \ldots, s_L)$ the strides of the L layers. We refer to the triple $(\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s})$ as the *architecture* of the LCN. We tacitly assume in the following that the dimensions $\boldsymbol{d}$ are compatible by satisfying (2.4). We also note that by (2.4) the function space $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is fully determined by fixing d_L, $\boldsymbol{k}$, and $\boldsymbol{s}$. As such, we may omit any of the indices $\boldsymbol{d}$, $\boldsymbol{k}$, and $\boldsymbol{s}$ when clear from the context. The function space $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is the image of the *parameterization map*

$$(2.5) \qquad \mu_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \colon \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_L} \longrightarrow \mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}; \quad (w_1, \ldots, w_L) \longmapsto \overline{W},$$

which takes L filters to the product of their associated convolutional matrices.

Remark 2.1. Although we use $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ to denote the function space of LCNs, the sequence of dimensions $\boldsymbol{d}$ does not play an essential role. More precisely, since the product of convolutional matrices is again a convolutional matrix (see Proposition 2.2), we can also define the parameterization map (2.5) at the level of filters $\mu_{\boldsymbol{k},\boldsymbol{s}} \colon \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_L} \to \mathbb{R}^k; \ (w_1, \ldots, w_L) \mapsto \overline{w}$, where $\overline{w}$ is the filter of the product convolutional matrix.

---

[1] The expression in (2.2) is actually the *cross-correlation* between x and w_l or equivalently the convolution of x with the *adjoint filter* $w_l^* = (w_{l,k_l-1}, \ldots, w_{l,0})$. We refer to it simply as a "convolution" following common machine learning terminology. Exchanging all filters with their adjoints would have no effect on our discussion.

2.1. Compositions of convolutional matrices. It is well known that the composition of convolution operations is again a convolution. More precisely, for two convolutional maps $\alpha_1 : \mathbb{R}^{d_0} \to \mathbb{R}^{d_1}$ and $\alpha_2 : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ with filters w_1 and w_2, respectively, and satisfying (2.4), the composition is given by

$$(2.6) \qquad \begin{aligned} (\alpha_2 \circ \alpha_1)(x)_i &= \sum_{j \in [k_2]} w_{2,j} \cdot \alpha_1(x)_{is_2+j} = \sum_{j \in [k_2]} w_{2,j} \cdot \sum_{\ell \in [k_1]} w_{1,\ell} \cdot x_{(is_2+j)s_1+\ell} \\ &= \sum_{j \in [k_2]} \sum_{\ell \in [k_1]} w_{2,j} w_{1,\ell} \cdot x_{is_2 s_1 + (js_1 + \ell)} = \sum_{m \in [(k_2-1)s_1+k_1]} u_m \cdot x_{is_2 s_1 + m}. \end{aligned}$$

Here the resulting filter u has size $(k_2 - 1)s_1 + k_1$ and has entries

$$(2.7) \qquad u_m = \sum_{\substack{j \in [k_2], \ \ell \in [k_1] \\ js_1 + \ell = m}} w_{2,j}\, w_{1,\ell}.$$

Proposition 2.2. *The composition of L convolutions with filter sizes $\boldsymbol{k} = (k_1, \ldots, k_L)$ and strides $\boldsymbol{s} = (s_1, \ldots, s_L)$ is a convolution with filter size $k = k_1 + \sum_{l=2}^{L}(k_l - 1)\prod_{m=1}^{l-1} s_m$ and stride $s = \prod_{l=1}^{L} s_l$. If all strides s_i are equal to $\tilde{s}$, then $k = d_0 - (d_L - 1)s$ and $s = \tilde{s}^L$.*

Proof. The expression in (2.6) shows that the composition of two convolutions with filter sizes k_1, k_2 and strides s_1, s_2 is again a convolution, with filter size $k = k_1 + (k_2 - 1)s_1$ and stride $s_2 s_1$. If the strides s_1, s_2 are equal to $\tilde{s}$, then by (2.4) we have $k_1 = d_0 - (d_1 - 1)\tilde{s}$ and $k_2 = d_1 - (d_2 - 1)\tilde{s}$, which yields $k = k_1 + (k_2 - 1)\tilde{s} = d_0 - (d_2 - 1)\tilde{s}^2$. The statement then follows by induction, whereby at the lth iteration we multiply the convolution represented up to layer $l - 1$ with the lth layer convolution. ∎

By Proposition 2.2, $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is contained in the vector space $\mathcal{M}_{(d_0,d_L),k,s}$ of convolutional $d_L \times d_0$ matrices with filter size k and stride s. A natural question is whether this containment is strict.

Definition 2.3. *An architecture $(\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s})$ is* filling *if $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} = \mathcal{M}_{(d_0,d_L),k,s}$ holds, where k and s are as in Proposition 2.2.*

Filling LCN architectures might be desirable since in function space training using a convex loss is simply convex optimization in a vector space. We emphasize that the vector space $\mathcal{M}_{(d_0,d_L),k,s}$ is properly contained in $\mathbb{R}^{d_L \times d_0}$ since convolutional matrices have restricted patterns of nonzero entries as well as rows with repeated entries. For nonfilling architectures, we will see in section 4 that $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ has additional constraints in the form of polynomial equalities and inequalities. Below we give examples of filling and nonfilling LCN architectures. For fully connected linear networks, the only constraints on the space of representable matrices are rank constraints (vanishing of certain minors, which are polynomial equations). For LCNs, the matrix sizes are decreasing by (2.4), and such rank constraints do not occur.

Example 2.4. Consider the LCN architecture with $\boldsymbol{d} = (3, 2, 1)$, $\boldsymbol{k} = (2, 2)$, $\boldsymbol{s} = (1, 1)$. The corresponding function space $\mathcal{M} \subset \mathbb{R}^{1 \times 3}$ consists of matrices

$$\overline{W} = W_2 W_1 = \begin{bmatrix} c & d \end{bmatrix} \begin{bmatrix} a & b & 0 \\ 0 & a & b \end{bmatrix} = \begin{bmatrix} ac & ad+bc & bd \end{bmatrix} =: \begin{bmatrix} A & B & C \end{bmatrix}.$$

We can see that this architecture is not filling, i.e., that the containment $\mathcal{M} \subsetneq \mathbb{R}^{1 \times 3}$ is strict. Indeed, the entries in $\overline{W}$ are given by the coefficients (A, B, C) of quadratic polynomials that factor into two real linear forms:

$$(a\mathsf{x} + b)(c\mathsf{x} + d) = ac\mathsf{x}^2 + (ad + bc)\mathsf{x} + bd = A\mathsf{x}^2 + B\mathsf{x} + C.$$

These correspond to quadratic equations with one or more real roots (or linear equations with $A = 0$). Hence the set of representable coefficients (A, B, C) is characterized by the discriminant condition $B^2 - 4AC \geq 0$. See Figure 1(left).

Example 2.5. Consider the LCN with $\boldsymbol{d} = (5, 3, 2)$, $\boldsymbol{k} = (3, 2)$, $\boldsymbol{s} = (1, 1)$. Then

$$\overline{W} = W_2 W_1 = \begin{bmatrix} d & e & 0 \\ 0 & d & e \end{bmatrix} \begin{bmatrix} a & b & c & 0 & 0 \\ 0 & a & b & c & 0 \\ 0 & 0 & a & b & c \end{bmatrix} = \begin{bmatrix} A & B & C & D & 0 \\ 0 & A & B & C & D \end{bmatrix},$$

where $(A, B, C, D) = (ad, bd + ae, cd + be, ce)$. The set $\mathcal{M}$ is the set of all possible $\overline{W}$ where its entries are related by the above equation as (a, b, c, d, e) vary over $\mathbb{R}^5$. Rephrasing this question as which real cubic polynomials factor into a real quadratic and linear factor, i.e.,

$$(a\mathsf{x}^2 + b\mathsf{x} + c)(d\mathsf{x} + e) = ad\mathsf{x}^3 + (bd + ae)\mathsf{x}^2 + (cd + be)\mathsf{x} + ce = A\mathsf{x}^3 + B\mathsf{x}^2 + C\mathsf{x} + D,$$

one can immediately draw on the fundamental theorem of algebra to conclude that every choice of (A, B, C, D) has a corresponding choice of (a, b, c, d, e). From this, we note that $\mathcal{M}$ is filling, so $\mathcal{M}_{(5,3,2),(3,2),(1,1)} = \mathcal{M}_{(5,2),4,1}$.

2.2. Polynomial multiplication. The previous examples illustrate a useful identification between convolutional matrices and polynomials. More precisely, we associate a filter $w \in \mathbb{R}^k$ with the polynomial

$$(2.8) \qquad \pi(w) = w_0 \mathsf{x}^{k-1} + w_1 \mathsf{x}^{k-2}\mathsf{y} + \cdots + w_{k-2}\mathsf{x}\mathsf{y}^{k-2} + w_{k-1}\mathsf{y}^{k-1} \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1}.$$

This is a formal polynomial, where the monomials $\mathsf{x}^i \mathsf{y}^j$ are used to index the weights of the filter. To better handle situations where leading coefficients vanish, it is convenient to use homogeneous bivariate polynomials instead of univariate affine polynomials as in the two previous examples. Recall that a root of a homogeneous polynomial $p(\mathsf{x}, \mathsf{y})$ is an equivalence class of pairs $(r_\mathsf{x}, r_\mathsf{y}) \in \mathbb{C}^2 \setminus \{(0,0)\}$ such that $p(r_\mathsf{x}, r_\mathsf{y}) = 0$, up to complex scalar factors (i.e., a point in $\mathbb{CP}^1$). With some abuse of notation, we will also use π to map a convolutional matrix W with filter w to the polynomial in (2.8).

Proposition 2.6. *Consider convolutional matrices $W_1, \ldots, W_L$ with compatible sizes and stride one. Then the convolutional matrix $\overline{W} = W_L \cdots W_1$ satisfies $\pi(\overline{W}) = \pi(W_L) \cdots \pi(W_1)$.*

Proof. For the product of two convolutional matrices, the statement follows from (2.7). The general case follows by induction. ∎

Corollary 2.7. *Let $(\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s})$ be an LCN architecture of depth L with $\boldsymbol{s} = (1, \ldots, 1)$. Then π identifies $\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}}$ with the set of polynomials that factor into polynomials of degrees $k_1 - 1, \ldots, k_L - 1$, that is, $\pi(\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}}) = \{p = q_L \cdots q_1 : q_i \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_{k_i - 1}\} \subseteq \mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1}$, where $k = \sum_{i=1}^{L} (k_i - 1) + 1$ is the size of the end-to-end filter.*

Since the ordering of the factors does not affect the product polynomial, the function space of an LCN with stride-one one-dimensional convolutional layers is the same for any permutation of the layers. This also holds for higher-dimensional convolutions, as we will see in subsection 4.3.

Remark 2.8. Proposition 2.6 can be generalized to deal with arbitrary strides $\boldsymbol{s}$. For example, for two convolutional matrices W_2, W_1 with strides s_2, s_1, the product $W_2 W_1$ can be identified with the polynomial $p = \pi(W_2 W_1) \in \mathbb{R}[\mathsf{x}, \mathsf{y}]$ that factors as $p = q_2 q_1$, where $q_1 = \pi(W_1)$ as in (2.8), while q_2 is defined as follows: if w denotes the filter of W_2, we see from (2.7) that

$$q_2 = w_0 \mathsf{x}^{s_1(k_2-1)} + w_1 \mathsf{x}^{s_1(k_2-2)} \mathsf{y}^{s_1} + \cdots + w_{k_2-2} \mathsf{x}^{s_1} \mathsf{y}^{s_1(k_2-2)} + w_{k_2-1} \mathsf{y}^{s_1(k_2-1)}.$$

2.3. Toeplitz vs. circulant matrices. Discrete convolutions over finite-dimensional data can be defined in slightly different ways depending on how they deal with the boundary of the input. Until now, we have been discussing a "no padding" setting where the convolution is not applied across the boundary, and the dimension of the output is smaller than the dimension of the input (see (2.3)). An alternative approach would be to consider *cyclic convolutions*, where the input space $\mathbb{R}^d$ is viewed as representing "cyclic signals" $\mathbb{Z}/d \to \mathbb{R}$. In this section, we point out that the two approaches are effectively equivalent in terms of the function space.

Let $W_1, \ldots, W_L$ be the weight matrices of an LCN. These matrices are generalized Toeplitz matrices and must have compatible sizes (2.4). If W_l has size $d_l \times d_{l-1}$ and corresponding filter w_l of size k_l, then we construct a *generalized circulant* matrix of size $d_0 \times d_0$ as follows. If the stride is $s_l = 1$, the circulant matrix is

$$\begin{bmatrix} w_{l,0} & w_{l,1} & \cdots & w_{l,d_0-1} \\ w_{l,d_0-1} & w_{l,0} & \cdots & w_{l,d_0-2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{l,1} & w_{l,2} & \cdots & w_{l,0} \end{bmatrix},$$

where $w_{l,i} = 0$ for $i \geq k_l$. For Toeplitz matrices with larger strides s_l, we can similarly define a $d_0 \times d_0$ generalized circulant matrix of stride s_l, where each row is obtained by shifting the previous row by s_l steps.

The generalized Toeplitz matrix W_l is the top-left $d_l \times d_{l-1}$ block of its associated generalized circulant matrix. Under this identification, the multiplication of generalized Toeplitz matrices with compatible sizes is the multiplication of the corresponding generalized circulant matrices. In particular, the resulting filter $\overline{w}$ of the Toeplitz matrix product $\overline{W}$ also results from the product of the circulant matrices. Therefore both formulations are equivalent.

Example 2.9. For the Toeplitz matrices in Example 2.4, the corresponding product of circulant matrices is

$$\begin{bmatrix} c & d & 0 \\ 0 & c & d \\ d & 0 & c \end{bmatrix} \begin{bmatrix} a & b & 0 \\ 0 & a & b \\ b & 0 & a \end{bmatrix} = \begin{bmatrix} ac & ad+bc & bd \\ bd & ac & ad+bc \\ ad+bc & bd & ac \end{bmatrix}.$$

3. Discriminants and multiple roots. In this section, we review some properties of the structure of real polynomials with repeated roots, recalling the definitions of *discriminants*, *multiple root loci*, and *real root multiplicity patterns* (*rrmp's* for short). We will use these notions in the next sections to study the function space of LCNs and to investigate the critical points of the associated loss functions. A reader who is more interested in the applications of these concepts to LCNs might skip this section on a first read and return to it later to better understand the structure of repeated roots in polynomials.

The *discriminant* of a polynomial $p = a_n\mathsf{x}^n + a_{n-1}\mathsf{x}^{n-1}\mathsf{y} + \cdots + a_1\mathsf{x}\mathsf{y}^{n-1} + a_0\mathsf{y}^n \in \mathbb{R}[\mathsf{x}, \mathsf{y}]$ is a polynomial in the coefficients $a_n, \ldots, a_0$ that vanishes if and only if p has a (complex) double root [10, Chapter 4]. The iterated singular loci of the discriminant hypersurface are the *coincident/multiple root loci*.

Definition 3.1. For a partition $\lambda = (\lambda_1, \ldots, \lambda_r)$ of $k - 1$, the multiple root locus $\Delta_\lambda \subset \mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1}$ *is the set of homogeneous polynomials that have r (not necessarily distinct) complex roots with multiplicities $\lambda_1, \ldots, \lambda_r$. That is, Δ_λ is the set of polynomials $P \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1}$ that can be written as $P = Q_1^{\lambda_1} \cdots Q_r^{\lambda_r}$, where each $Q_i \in \mathbb{C}[\mathsf{x}, \mathsf{y}]_1$ is a complex linear form.*

For instance, the discriminant hypersurface is $\Delta_{(2,1,\ldots,1)}$. The singular locus of a multiple root locus is a union of certain higher-order multiple root loci; see [25] for a combinatorial description.[2] We are also interested in distinguishing between real and complex roots. This more fine-grained information is captured by the *rrmp*.

Definition 3.2. A homogeneous polynomial $P \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1}$ has the real root multiplicity pattern*, rrmp for short, $(\rho \mid \gamma) = (\rho_1, \ldots, \rho_r \mid \gamma_1, \ldots, \gamma_c)$ if it can be written as*

$$(3.1) \qquad P = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c},$$

where $p_i \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_1$ and $q_j \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_2$ are irreducible and pairwise linearly independent.

Recall that each factor q_j has two complex conjugate roots. Polynomials with rrmp $(\rho \mid \gamma)$ are smooth points on the multiple root locus $\Delta_{\lambda_{\rho|\gamma}}$, where $\lambda_{\rho|\gamma} = (\rho_1, \ldots, \rho_r, \gamma_1, \gamma_1, \ldots, \gamma_c, \gamma_c)$.

Example 3.3. We describe the possible rrmp of a quadratic, cubic, and quartic polynomial. The discriminant polynomial Δ of $a\mathsf{x}^2 + b\mathsf{x}\mathsf{y} + c\mathsf{y}^2$ distinguishes three rrmp's:

11|0	2 distinct real roots:	$\Delta = b^2 - 4ac > 0$.
2|0	1 double real root:	$\Delta = 0$.
0|1	no real roots:	$\Delta < 0$.

For cubic polynomials $a\mathsf{x}^3 + b\mathsf{x}^2\mathsf{y} + c\mathsf{x}\mathsf{y}^2 + d\mathsf{y}^3$ with real coefficients, there are four rrmp's shown below, discussed in [3, 11, 18].

111|0	3 distinct real roots:	$\Delta = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd > 0$.
12|0	1 single real root and 1 double real root:	$\Delta = 0$.
3|0	1 triple real root:	$\delta_1 = 3ac - b^2 = 0$ and $\delta_2 = 9ad - bc = 0$ and $\delta_3 = 3bd - c^2 = 0$.
1|1	1 real root and 2 complex conjugate roots:	$\Delta < 0$.

---

[2]For two partitions λ and μ of $k - 1$, Δ_μ is contained in Δ_λ if and only if μ is a coarsening of λ. In [25, Proposition 2.1(i)], it is shown that Δ_μ is contained in the singular locus of Δ_λ if and only if either there are at least two ways of obtaining λ as a splitting of μ or there is a splitting such that one of the elements μ_i in μ is split into subsets which do not all have the same size. An illustrative example is given in [25, Example 2.4].

Following [3], a polynomial $f_n = \mathsf{x}^n + a_1\mathsf{x}^{n-1}\mathsf{y} + a_2\mathsf{x}^{n-2}\mathsf{y}^2 + \cdots + a_n\mathsf{y}^n$ can be translated to a polynomial $g_n = f_n(\mathsf{x} - {}^{a_1}/n\mathsf{y}, \mathsf{y}) = \mathsf{x}^n + b_2\mathsf{x}^{n-2}\mathsf{y}^2 + b_3\mathsf{x}^{n-3}\mathsf{y}^3 + \cdots + b_n\mathsf{y}^n$. The two polynomials f_n and g_n have the same rrmp. For a quartic polynomial $g_4 = \mathsf{x}^4 + p\mathsf{x}^2\mathsf{y}^2 + q\mathsf{x}\mathsf{y}^3 + r\mathsf{y}^4$, letting

$$\delta(p, q, r) = 256r^3 - 128p^2r^2 + 144pq^2r + 16p^4r - 27q^4 - 4p^3q^2,$$
$$\delta'(p, q, r) = 8pr - 9q^2 - 2p^3,$$

where δ corresponds to the discriminant, the different rrmp's are as follows:

1111|0 4 distinct real roots: $\delta > 0$ and $\delta' > 0$ and $p < 0$.
112|0 2 distinct real roots and 1 double real root: $\delta = 0$ and $\delta' > 0$ and $p < 0$.
22|0 2 real double roots: $\delta = 0$ and $\delta' = 0$ and $p < 0$ and $q = 0$.
13|0 1 single real root and 1 triple real root: $\delta = 0$ and $\delta' = 0$ and $p < 0$ and $q \neq 0$.
4|0 1 quadruple real root: $\delta = 0$ and $\delta' = 0$ and $p = 0$.
11|1 2 distinct real roots and 2 distinct complex conjugate roots: $\delta < 0$.
2|1 1 double real root and 2 distinct complex conjugate roots: $\delta = 0$ and $\delta' < 0$.
0|2 2 complex conjugate double roots: $\delta = 0$ and $\delta' = 0$ and $p > 0$.
0|11 4 distinct complex roots: $\delta > 0$ and ($\delta' \leq 0$ or $p > 0$).

4. Function space of linear convolutional networks. We show that the function space of an LCN is a semi-algebraic set and discuss its dimension and whether it is filling (Definition 2.3).

4.1. Convolutional matrices of stride one. We first focus on one-dimensional convolutions and stride one. Recall that a semi-algebraic set is a solution set to a finite set of polynomial equations and polynomial inequalities, or a finite union of such sets.

Theorem 4.1. *Let $(\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s})$ be an LCN architecture with $\boldsymbol{s} = (1, \ldots, 1)$ and depth L. Then $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is a full-dimensional semi-algebraic subset of $\mathcal{M}_{(d_0,d_L),k,s}$, where $k = \sum_{i=1}^{L} k_i - L + 1$ and $s = 1$. In particular, $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is cut out from $\mathcal{M}_{(d_0,d_L),k,s}$ by inequalities only. Moreover, the architecture is filling if and only if at most one of the filter sizes $k_1, \ldots, k_L$ is even.*

This result explains the difference between Examples 2.4 and 2.5. To prove it, we use Corollary 2.7: We identify $\mathcal{M}_{(d_0,d_L),k,s}$ with the space $\mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1}$ of polynomials of degree $k-1$ and the subset $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ with the polynomials that factor as a product of polynomials of degrees $k_1 - 1, \ldots, k_L - 1$. Under this identification, we have the following.

Lemma 4.2. *Let $\boldsymbol{s} = (1, \ldots, 1)$. A polynomial $p \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1}$ corresponds to a convolutional matrix in $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ if and only if p has at least $e := |\{k_i : k_i \text{ is even}\}|$ real roots (counted with multiplicity), i.e., if the rrmp $(\rho \mid \gamma)$ of p satisfies $\sum_{i=1}^{r} \rho_i \geq e$.*

Proof. By Corollary 2.7, it suffices to show that a polynomial p of degree $k - 1$ has at least e real roots if and only if p is a product of polynomials of degrees $k_1 - 1, \ldots, k_L - 1$. One direction is immediate. A factorization of p of the desired type has e odd-degree factors, and any polynomial p with e odd-degree factors has at least e real roots.

For the other direction, let e' denote the number of real roots of p. The irreducible factors of p are e' linear terms and $(k - 1 - e')/2$ quadratic terms. If $e' \geq e$, these factors can be multiplied to build a set of factors $q_1, \ldots, q_L$ of degrees $k_1 - 1, \ldots, k_L - 1$ as follows. We first assign one of the e' linear terms to each of the e factors q_i of odd degree, and then we use the

remaining $(e' - e) + (k - 1 - e')/2$ irreducible factors of p to fill up the factors $q_1, \ldots, q_L$ until their degrees are indeed $k_1 - 1, \ldots, k_L - 1$. Note that this procedure works as the degrees of the irreducible factors of p and the degrees of $q_1, \ldots, q_L$ both sum up to the same value $k - 1$. ∎

Example 4.3. When $\boldsymbol{s} = (1, \ldots, 1)$ and $\boldsymbol{k} = (k_1, \ldots, k_L)$ are chosen such that the resulting filter size is $k = 5$, by Lemma 4.2 the matrices in $\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}}$ correspond to quartic polynomials with a particular real root structure, specified by the individual filter sizes k_i:

1. If $L = 2$ and $(k_1, k_2) = (3, 3)$, then every polynomial of degree four is in $\pi(\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}})$.
2. If $L = 2$ and $(k_1, k_2) = (4, 2)$, then a quartic polynomial p corresponds to a matrix in the function space if and only if it has two or four real roots. Hence, the complement of the function space corresponds to all quartics with no real roots.
3. If $L = 3$ and $(k_1, k_2, k_3) = (3, 2, 2)$, this architecture has the same function space as the previous one. However, the parameterization of the function space is different: Here $\pi(\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}})$ arises by taking the product of two linear factors and one quadratic factor, whereas in the previous architecture one considers the product of a linear and a cubic factor.
4. If $L = 4$ and $(k_1, k_2, k_3, k_4) = (2, 2, 2, 2)$, a polynomial of degree four is in $\pi(\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}})$ if and only if all its roots are real.

Proof of Theorem 4.1. We first consider the case where all filter sizes are odd. Using Lemma 4.2, we see that $e = 0$, so every polynomial of degree $k - 1$ corresponds to a matrix in the function space $\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}}$. If exactly one filter size is even, then $k - 1$ is odd. Hence, every polynomial p of degree $k - 1$ has at least one real root. In Lemma 4.2, we have $e = 1$, which shows that $p \in \pi(\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}})$. We conclude that the architecture is filling if at most one filter size is even. Finally, assume that two or more filter sizes are even. In this case, we consider a polynomial p of degree $k - 1$ that has zero (if $k - 1$ is even) or one (if $k - 1$ is odd) real root. Then p does not belong to $\pi(\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}})$ by Lemma 4.2. This shows that the architecture is not filling. That the function space is semi-algebraic follows from the fact that it has a polynomial parameterization (by Tarski–Seidenberg). Finally, we argue that the set of polynomials representable by a given LCN is full dimensional in $\mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1}$. For this, it is enough to notice that small perturbations of the coefficients of a generic polynomial will not affect its root structure. We will discuss this in more detail in the following statements. ∎

When the architecture of an LCN is not filling, it is interesting to study the boundary of the function space.

Proposition 4.4. *Let $(\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s})$ be an LCN architecture with $e := |\{k_i \colon k_i \text{ is even}\}| \geq 2$ and $\boldsymbol{s} = (1, \ldots, 1)$. In the Euclidean topology on $\mathcal{M}_{(d_0, d_L), \boldsymbol{k}, \boldsymbol{s}}$, the function space $\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}}$ is closed and its boundary consists of all convolutional matrices that correspond to polynomials with rrmp $(\rho \mid \gamma)$ satisfying $\sum_{i=1}^{r} \rho_i \geq e$ and $|\{\rho_i \colon \rho_i \text{ is odd}\}| \leq e - 2$.*

Proof. A sequence of polynomials, each having at least e real roots counted with multiplicity, will also have at least e real roots in the limit. Hence, by Lemma 4.2, the function space is closed. Using the identification π with polynomials, the boundary of $\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}}$ consists of all polynomials in $\mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}}$ that are limits of sequences of polynomials in its complement.

Given a sequence of polynomials $P^{(j)}$ of degree $(k - 1) = \sum_{i=1}^{L} (k_i - 1)$ in the complement

of $\mathcal{M}_{d,k,s}$ that converges to a polynomial in $\mathcal{M}_{d,k,s}$, we may assume (by restricting to a subsequence) that all $P^{(j)}$ have the same rrmp $(\rho \mid \gamma)$. Since e and $(k-1)$ have the same parity and $P^{(j)} \notin \mathcal{M}_{d,k,s}$, we have $\sum_{i=1}^{r} \rho_i \leq e - 2$. In particular, the number of odd ρ_i is at most $e - 2$. In the limit of the $P^{(j)}$, pairs of complex conjugate roots might become real double roots (which does not affect the number of odd ρ_i) or distinct real roots might become equal (which cannot increase the number of odd ρ_i). Hence, the limit of the $P^{(j)}$ has rrmp $(\rho' \mid \gamma')$ with at most $e - 2$ odd ρ'_i and $\sum \rho'_i \geq e$ (since we assumed the limit to be in $\mathcal{M}_{d,k,s}$).

Conversely, given a polynomial P with rrmp $(\rho \mid \gamma)$ satisfying $|\{\rho_i \colon \rho_i \text{ is odd}\}| \leq e - 2$ and $\sum_{i=1}^{r} \rho_i \geq e$, we write $\rho_i = 2\alpha_i + \beta_i$, where $\beta_i \in \{0, 1\}$. We consider its factorization as in (3.1) and perturb each factor $p_i^{\rho_i}$ to a polynomial with α_i pairs of complex conjugate roots, plus one real root if $\beta_i = 1$. That way, P gets perturbed to a polynomial $\tilde{P}$ with exactly $\sum \beta_i$ real roots. We have that $\sum \beta_i = |\{\rho_i \colon \rho_i \text{ is odd}\}| \leq e - 2$, so $\tilde{P} \notin \mathcal{M}_{d,k,s}$. ∎

Corollary 4.5. *Let (d, k, s) be an LCN architecture with $s = (1, \ldots, 1)$ and two or more even filter sizes. The Zariski closure of the Euclidean boundary of the function space $\mathcal{M}_{d,k,s}$ is the discriminant hypersurface $\pi^{-1}(\Delta_{(2,1,\ldots,1)})$.*

Proof. Let $e := |\{k_i \colon k_i \text{ is even}\}|$. Since the rrmp $(\rho \mid \gamma)$ of any polynomial on the Euclidean boundary of $\mathcal{M}_{d,k,s}$ satisfies by Proposition 4.4 that at least one of the ρ_i is larger than one, the boundary is contained in the discriminant hypersurface. Moreover, any polynomial with one real double root, $e - 2$ pairwise distinct real single roots, and all other roots nonreal with multiplicity one is both on the boundary of $\mathcal{M}_{d,k,s}$ and a smooth point of the discriminant. ∎

Example 4.6. The following table lists all nonfilling architectures with stride one and where the end-to-end filter has sizes $3, 4, 5$ (up to permutations of the filters):

k	function space $\mathcal{M}_k$	complement	Euclidean boundary
$(2,2)$	$11\|0,\ 2\|0$	$0\|1$	$2\|0$
$(2,2,2)$	$111\|0,\ 12\|0,\ 3\|0$	$1\|1$	$12\|0,\ 3\|0$
$(3,2,2)$ or $(4,2)$	$1111\|0,\ 112\|0,\ 22\|0,\ 13\|0,\ 4\|0,\ 11\|1,\ 2\|1$	$0\|2,\ 0\|11$	$2\|1,\ 22\|0,\ 4\|0$
$(2,2,2,2)$	$1111\|0,\ 112\|0,\ 22\|0,\ 13\|0,\ 4\|0$	$11\|1,\ 2\|1,\ 0\|2,\ 0\|11$	$112\|0,\ 13\|0,\ 22\|0,\ 4\|0$

For each architecture k, the function space $\mathcal{M}_k$ consists of all polynomials with one of the rrmp's listed above. We described the possible rrmp's for quadratic, cubic, and quartic polynomials in Example 3.3. Similarly, the complement and boundary of $\mathcal{M}_k$ are as shown above.

When exactly two filter sizes are even, the nonfilling geometry of $\mathcal{M}_{d,k,s}$ has a particularly simple description.

Proposition 4.7. *Let (d, k, s) be an LCN architecture with $s = (1, \ldots, 1)$ and exactly two even filter sizes. The complement of $\pi(\mathcal{M}_{d,k,s})$ in $\mathbb{R}[x, y]_{k-1}$ is a union of two open convex cones, each consisting of all positive or all negative polynomials.*

Proof. If exactly two filters have even size, then $k - 1$ is even, and a polynomial of degree $k - 1$ is in $\pi(\mathcal{M}_{d,k,s})$ if and only if it has at least two real roots (by Lemma 4.2). Hence, the complement of $\pi(\mathcal{M}_{d,k,s})$ consists of all polynomials of even degree $k - 1$ without real roots. ∎

This behavior is exhibited in Example 2.4 and illustrated on the left of Figure 1 and in

Example 4.3, cases 2 and 3 (see also Example 4.6, first and third rows). In general, if strictly more than two filter sizes are even, the complement of $\pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}})$ is *not* a union of two convex cones, as the following example demonstrates.

Example 4.8. Consider the case $|\{k_i \colon k_i \text{ is even}\}| = 3$ with stride one. A polynomial of odd degree $k - 1$ is in $\pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}})$ if and only if p has at least three real roots (by Lemma 4.2). Hence, the complement $\mathbb{R}[\mathsf{x},\mathsf{y}]_{k-1} \setminus \pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}})$ is the set of all polynomials of degree $k-1$ with exactly one real root. Even when restricting ourselves to degree-$(k-1)$ polynomials with one real root such that the remaining degree-$(k-2)$ factor is positive, this set does not form a convex cone. For instance, if $k - 1 = 3$, consider the polynomials $(\mathsf{x} - \mathsf{y})\mathsf{x}(\mathsf{x} + \mathsf{y}) + 2\mathsf{y}^3$ and $(\mathsf{x} - \mathsf{y})\mathsf{x}(\mathsf{x} + \mathsf{y}) - 2\mathsf{y}^3$: Both have exactly one real root such that the remaining quadratic factor is positive. However, their sum clearly has three real roots.

4.2. Generalization to larger strides. In this section, we consider one-dimensional convolutions that may have stride larger than one. In this setting, the identification with polynomials holds as described in Remark 2.8. Specifically, we write

$$\pi_s : \mathbb{R}^k \longrightarrow \mathbb{R}[\mathsf{x},\mathsf{y}]_{(k-1)s}, \quad w \longmapsto w_0 \mathsf{x}^{(k-1)s} + w_1 \mathsf{x}^{(k-2)s}\mathsf{y}^s + \cdots + w_{k-2}\mathsf{x}^s \mathsf{y}^{(k-2)s} + w_{k-1}\mathsf{y}^{(k-1)s}.$$

This generalizes the map $\pi = \pi_1$ defined in (2.8). If W_2 and W_1 are convolutional matrices with strides s_2 and s_1, then $\pi(W_2 W_1) = \pi_{s_1}(W_2)\pi(W_1)$. Note that π_{s_1} is applied to W_2, while s_2 does not have any effect on the filter of $W_2 W_1$ (only on its stride, which is given by $s_2 s_1$). Based on this fact, we deduce the following two results.

Lemma 4.9.(a) *If* $\boldsymbol{s} = (s_1, \ldots, s_L)$ *and* $\boldsymbol{s}' = (s_1, \ldots, s_{L-1}, 1)$, *then* $\pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}) = \pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}'})$.
(b) *If* $\boldsymbol{k} = (k_1, \ldots, k_{L-1}, 1)$ *and* $\boldsymbol{k}' = (k_1, \ldots, k_{L-1})$, $\boldsymbol{d}' = (d_0, \ldots, d_{L-1})$, $\boldsymbol{s}' = (s_1, \ldots, s_{L-1})$, *then* $\pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}) = \pi(\mathcal{M}_{\boldsymbol{d}',\boldsymbol{k}',\boldsymbol{s}'})$.

Proof. This follows from the fact that $\pi(W_L \cdots W_1) = \pi_s(W_L)\pi_1(W_{L-1} \cdots W_1)$, where $s = s_{L-1} \ldots s_1$ is the stride of $W_{L-1} \cdots W_1$. Part (a) holds since this fact is independent of s_L. Part (b) holds since $k_L = 1$ implies that $\pi_s(W_L)$ is a degree-0 polynomial, which does not play a role in the factorization of the end-to-end filter. ∎

Corollary 4.10. *Let* $s_1 = \cdots = s_{L-1} = 1$ *and* s_L *be arbitrary. Then the LCN architecture* $(\boldsymbol{d},\boldsymbol{k},\boldsymbol{s})$ *satisfies the properties in Theorem* 4.1 *(with* $s = s_L$*) and Propositions* 4.4 *and* 4.7.

The corollary is immediately implied by Lemma 4.9(a). This covers the case where only the last layer has an arbitrary stride. Similarly, Lemma 4.9(b) says that if the last filter size is one, we can understand the function space from the smaller architecture that only consists of the first $L - 1$ layers. For the remaining cases, we obtain the following result.

Proposition 4.11. *If* $k_L > 1$ *and* $s_i > 1$ *for some* $i \leq L - 1$, *the LCN architecture is non-filling. In fact, the function space is a lower-dimensional semi-algebraic subset of* $\mathcal{M}_{(d_0,d_L),\boldsymbol{k},\boldsymbol{s}}$.

Proof. Consider the last product $W_L \cdot W'$ where $W' = W_{L-1} \cdots W_1$. We write $k' = k_1 + \sum_{l=2}^{L-1}(k_l - 1)\prod_{m=1}^{l-1} s_m$ for the filter size of W' and $s' = s_1 \cdots s_{L-1} > 1$ for its stride. Even if W' is an arbitrary convolutional matrix with filter size and stride (k', s') (i.e., even if the LCN architecture from the first $L - 1$ layers is filling), the product $W_L \cdot W'$ is *not* an *arbitrary* convolutional matrix with filter size $k = k' + (k_L - 1)s'$ and stride $s = s' s_L$. Otherwise, every polynomial $p \in \mathbb{R}[\mathsf{x},\mathsf{y}]_{k-1}$ could be factored into a real polynomial p' of

degree $k' - 1$ in x, y and a real polynomial p_L of degree $k_L - 1$ in $\mathsf{x}^{s'}, \mathsf{y}^{s'}$ by Remark 2.8. Modding out global scaling, we may assume that the polynomial p and its two factors p' and p_L are monic, so their remaining degrees of freedom (i.e., the number of their remaining coefficients) are $k - 1$ for p, $k' - 1$ for p', and $k_L - 1$ for p_L. Since $k_L > 1$ and $s' > 1$, we have that $(k' - 1) + (k_L - 1) < k' + (k_L - 1)s' - 1 = k - 1$, so the monic polynomials p' and p_L do not have enough degrees of freedom for their product to yield all monic polynomials p. This dimension count shows that $\dim(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}) \leq k' + k_L - 1 < k = \dim(\mathcal{M}_{(d_0,d_L),k,s})$. ∎

For the nonfilling architectures in Proposition 4.11, $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ has a strictly smaller dimension than $\mathcal{M}_{(d_L,d_0),k,s}$ and is thus cut out by polynomial equations and inequalities. This is in contrast to the nonfilling architectures appearing for stride one in Theorem 4.1: in that case, $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ has the same dimension as $\mathcal{M}_{(d_0,d_L),k,s}$ and is cut out only by polynomial inequalities.

Example 4.12. Consider the LCN with $\boldsymbol{d} = (5,2,1)$, $\boldsymbol{k} = (3,2)$, $\boldsymbol{s} = (2,1)$. Note $s_2 = 1$ is inconsequential by Lemma 4.9(a). This setting is similar to Example 2.5, where we considered $\boldsymbol{s} = (1,1)$. In that case, the architecture was filling, but this is no longer the case for stride size 2, as we will see. In the current example, the function space $\mathcal{M} \in \mathbb{R}^{1 \times 5}$ consists of matrices

$$(4.1) \qquad \overline{W} = W_2 W_1 = \begin{bmatrix} d & e \end{bmatrix} \begin{bmatrix} a & b & c & 0 & 0 \\ 0 & 0 & a & b & c \end{bmatrix} = \begin{bmatrix} ad & bd & ae + cd & be & ce \end{bmatrix}.$$

We see that the coefficients in (4.1) are given by

$$(4.2) \qquad \pi(\overline{W}) = (d\mathsf{x}^2 + e\mathsf{y}^2)(a\mathsf{x}^2 + b\mathsf{x}\mathsf{y} + c\mathsf{y}^2).$$

The architecture is filling if and only if every real quartic polynomial $A\mathsf{x}^4 + B\mathsf{x}^3\mathsf{y} + C\mathsf{x}^2\mathsf{y}^2 + D\mathsf{x}\mathsf{y}^3 + E\mathsf{y}^4$ can be factored into two real quadratic polynomials as above. However, it is easy to verify that the coefficients of the polynomial in (4.2) satisfy the equation $AD^2 + B^2E = BCD$. In fact, we can fully characterize the function space as follows:

$$(4.3) \qquad \mathcal{M} = \{(A,B,C,D,E) \in \mathbb{R}^5 \colon AD^2 + B^2E = BCD \text{ and } C^2 \geq 4AE\}.$$

Indeed, the coefficients in (4.1) satisfy $(ae + cd)^2 - 4adce = (ae - cd)^2 \geq 0$. For the other containment, we consider a tuple $(A, \ldots, E)$ in the right-hand side of (4.3) and explicitly find a factorization as in (4.2). For the case $E = D = 0$, set $a = A$, $b = B$, $c = C$, $d = 1$, $e = 0$. For the case $E = 0, D \neq 0$, set $a = C/D$, $b = 1$, $c = 0$, $d = B$, $e = D$, which works since $AD = BC$. It remains to show the case $E \neq 0$. We may assume without loss of generality that $E = 1$, and we set $c = 1, e = 1, b = D$. To determine a and d, we distinguish four cases:

- $D \neq 0, B \neq 0$: Set $d = B/D$ and $a = AD/B$.
- $D \neq 0, B = 0$: Set $d = 0$ and $a = C$.
- $D = 0, A \neq 0$: At least one of the two values $\frac{C}{2} \pm \sqrt{\frac{C^2}{4} - A}$ is nonzero since $A \neq 0$. Set d to be such a nonzero value and $a = A/d$.
- $D = 0, A = 0$: Set $a = C$ and $d = 0$.

Using the equation and inequality in the right-hand side of (4.3), we verify that these four cases provide factorizations as in (4.2).

4.3. Higher-dimensional convolutions. We briefly discuss convolutions over inputs having a D-dimensional structure, $D > 1$ (e.g., raster images). The output of each layer is now a tensor of order D. Thus, we identify $\mathbb{R}^{d_l} \cong \mathbb{R}^{\Lambda_l}$ with $\Lambda_l = [d_l^1] \times \cdots \times [d_l^D]$. Each filter is also a tensor $w_l \in \mathbb{R}^{\lambda_l}$ with $\lambda_l = [k_l^1] \times \cdots \times [k_l^D]$. The convolution of $x \in \mathbb{R}^{\Lambda_{l-1}}$ with $w_l \in \mathbb{R}^{\lambda_l}$ is

$$(4.4) \qquad (\alpha_l(x))_i = \sum_{j \in \lambda_l} w_{l,j} \cdot x_{is_l+j}, \quad i \in \Lambda_l,$$

where $s_l \in \mathbb{Z}_{>0}$ is the stride of the convolution (cf. (2.2) for one-dimensional convolutions). The map (4.4) is linear in x and can be represented as a *convolutional tensor* $T_l \in \mathbb{R}^{\Lambda_l \times \Lambda_{l-1}}$ of order 2D. The dimensions of the tensor T_l satisfy (2.4) along each dimension: $d_l^h = \frac{d_{l-1}^h - k_l^h}{s_l} + 1$.

Example 4.13. Let $D = 2$, and let us consider a filter w of size 2×2 with stride one. The convolution of an input $x = \begin{bmatrix} * & * \\ *\circ & *\circ \\ \circ & \circ \end{bmatrix}$ of size 3×2 with the filter w yields an output of size 2×1: Its first (resp., last) entry is the inner product of w with the first (resp., last) four entries of x marked by $*$ (resp., $\circ$). The associated convolutional tensor T has size $2 \times 1 \times 3 \times 2$, and its entries (shown in two slices) are as follows:

$$T_{00::} = \begin{bmatrix} w_{00} & w_{01} \\ w_{10} & w_{11} \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad T_{10::} = \begin{bmatrix} 0 & 0 \\ w_{00} & w_{01} \\ w_{10} & w_{11} \end{bmatrix}.$$

In the following, we assume that all strides are one. Similar to the one-dimensional case, we can represent convolutions using polynomials. If $w \in \mathbb{R}^{\lambda_l}$, we define

$$(4.5) \qquad \pi(w) = \sum_{i \in \lambda_l} w_{i_1,\ldots,i_D}\, \mathsf{x}_1^{i_1} \mathsf{y}_1^{k_l^1-1-i_1} \cdots \mathsf{x}_D^{i_D} \mathsf{y}_D^{k_l^D-1-i_D} \ \in \ \mathbb{R}[\mathsf{x}_1,\mathsf{y}_1,\ldots,\mathsf{x}_D,\mathsf{y}_D],$$

where the weights of the filter are the coefficients of a multivariate polynomial that is multi-homogeneous, that is, it has degree $k_l^j - 1$ in each pair $\mathsf{x}_j, \mathsf{y}_j$. For instance, in Example 4.13 we have $\pi(w) = w_{00}\mathsf{y}_1\mathsf{y}_2 + w_{10}\mathsf{x}_1\mathsf{y}_2 + w_{01}\mathsf{y}_1\mathsf{x}_2 + w_{11}\mathsf{x}_1\mathsf{x}_2$. If T is a convolutional tensor with filter w, we also write $\pi(T)$ for the polynomial in (4.5). The following is a direct generalization of Proposition 2.6.

Proposition 4.14. *Consider a collection of convolutional tensors $T_1, \ldots, T_L$ with compatible sizes and stride one. The tensor $\overline{T} = T_L \circ \cdots \circ T_1$ corresponding to the composition of the associated linear maps (4.4) is a convolutional tensor and satisfies $\pi(\overline{T}) = \pi(T_L) \cdots \pi(T_1)$.*

Proof. The proof follows from the composition of two maps as in (4.4) with stride one:

$$(\alpha_2 \circ \alpha_1)(x)_i = \sum_{j \in \lambda_2} w_{2,j} \cdot \alpha_1(x)_{i+j} = \sum_{j \in \lambda_2} w_{2,j} \cdot \sum_{\ell \in \lambda_1} w_{1,\ell} x_{(i+j)+\ell}$$

$$= \sum_{j \in \lambda_2} \sum_{\ell \in \lambda_1} w_{2,j} w_{1,\ell} x_{i+(j+\ell)} = \sum_{m \in \mu} u_m \cdot x_{i+m},$$

where $\mu = [k_1^1 + k_2^1 - 1] \times \cdots \times [k_1^D + k_2^D - 1]$ and $u \in \mathbb{R}^\mu$ is the filter with entries $u_m = \sum_{j \in \lambda_1, \ell \in \lambda_2, j+\ell=m} w_{2,j} w_{1,\ell}$. From this expression, it follows that $\pi(u) = \pi(w_2)\pi(w_1)$. $\blacksquare$

Corollary 4.15. *Given an LCN architecture with L layers, inputs of format $[d_0^1] \times \cdots \times [d_0^D]$, and stride one, we write $(k_i^1, \ldots, k_i^D)$ for the filter size in the ith layer. The function space can be identified with the family of polynomials $P \in \mathbb{R}[\mathsf{x}_1, \mathsf{y}_1, \ldots, \mathsf{x}_D, \mathsf{y}_D]$ that can be factored as $P = Q_L \cdots Q_1$, where Q_i is homogeneous of degree $k_i^j - 1$ in the variables $\mathsf{x}_j, \mathsf{y}_j$.*

The natural ambient space of the LCN function space is the set of polynomials $P \in \mathbb{R}[\mathsf{x}_1, \mathsf{y}_1, \ldots, \mathsf{x}_D, \mathsf{y}_D]$ that are homogeneous of degree $k^j - 1$ in $\mathsf{x}_j, \mathsf{y}_j$ where $k^j := (k_1^j - 1) + \cdots + (k_L^j - 1) + 1$. We denote this vector space by $\mathcal{M}_{(d_0, d_L), k, s}$, where $d_0 = (d_0^1, \ldots, d_0^D)$, $d_L = (d_L^1, \ldots, d_L^D)$, $k = (k^1, \ldots, k^D)$, $s = (1, \ldots, 1)$. Using Corollary 4.15, we show that for higher-dimensional convolutions and at least two layers (and excluding trivial cases where filters have size one) the LCN function space is always a lower-dimensional semi-algebraic subset of $\mathcal{M}_{(d_0, d_L), k, s}$ and in particular is not filling.

Corollary 4.16. *Consider an LCN architecture with stride one, $D > 1$, and $L > 1$, where no layer has trivial filter size $(1, \ldots, 1)$. If there are distinct $j_1, j_2 \in \{1, \ldots, D\}$ such that $k^{j_1} > 1$ and $k^{j_2} > 1$, the architecture is not filling. In fact, the function space is a lower-dimensional semi-algebraic subset of $\mathcal{M}_{(d_0, d_L), k, s}$.*

Proof. The statement is a consequence of the fact that multivariate polynomials are generically irreducible. Since at least two layers have nontrivial filter sizes, the function space is contained in the set of polynomials that are reducible over $\mathbb{C}$, which is Zariski closed and thus lower-dimensional in the ambient space $\mathcal{M}_{(d_0, d_L), k, s}$. As before, the fact that the function space is semi-algebraic follows from Tarski–Seidenberg.

For completeness, we give a formal argument showing that a generic polynomial P in the ambient space $\mathcal{M}_{(d_0, d_L), k, s}$ is irreducible. We may assume without loss of generality that $k^j > 1$ for all $1 \le j \le n$ and that $k^j = 1$ for all $n < j \le D$. The latter means that the variables $\mathsf{x}_j, \mathsf{y}_j$ for $n < j \le D$ do not appear in the polynomial P; see (4.5). Hence, the zero locus of P is a hypersurface in $(\mathbb{P}^1)^n$ of multidegree $(k^1 - 1, \ldots, k^n - 1)$. Given the assumptions of the architecture, note that $n \ge 2$. To proceed, we consider the Segre–Veronese embedding $(\mathbb{P}^1)^n \hookrightarrow \mathbb{P}^{N-1}$ that maps $(\mathsf{x}_j : \mathsf{y}_j)_{j=1}^n$ to the N-tuple of all monomials that are homogeneous of degree $k^j - 1$ in each pair $\mathsf{x}_j, \mathsf{y}_j$. Under this map, the polynomial P becomes linear in the new coordinates of $\mathbb{P}^{N-1}$. In turn, the zero locus of P is a hyperplane section of $(\mathbb{P}^1)^n$ embedded in $\mathbb{P}^{N-1}$. Now, Bertini's theorem [22, Chapter II, Theorem 8.18 and Chapter III, Remark 7.9.1] states that, for smooth complex projective varieties of dimension at least two (in our case, $(\mathbb{P}^1)^n$ with $n \ge 2$), generic hyperplane sections are irreducible. This implies that a generic polynomial P in $\mathcal{M}_{(d_0, d_L), k, s}$ is irreducible over $\mathbb{C}$. ∎

Example 4.17. Consider an LCN architecture with 3×3 matrices as inputs (i.e., $D = 2$ and $d_0 = (3, 3)$), $L = 2$ layers, filter sizes $(k_1^1, k_1^2) = (2, 2) = (k_2^1, k_2^2)$, and strides one. The two convolutions are maps $\mathbb{R}^{3 \times 3} \to \mathbb{R}^{2 \times 2} \to \mathbb{R}^{1 \times 1}$. Their composition is a convolution with filter size 3×3. Hence, the natural ambient space of the LCN function space is the nine-dimensional vector space of polynomials in two pairs of variables such that the degree in each pair is two, and the function space is the subset of polynomials that factor into two polynomials that are linear in each pair. The function space has codimension two in the ambient space. A `Macaulay2` [19] computation reveals that its Zariski closure has degree 10 and that its vanishing ideal is minimally generated by 39 homogeneous polynomials of degree 6.

5. Optimization. We study the optimization of an objective function, or *loss*, using LCNs with one-dimensional convolutions and stride one. A loss $\mathcal{L}(\theta)$ is a function of the filters $\theta = (w_1, \ldots, w_L)$ that can be written as a composition $\mathcal{L} = \ell \circ \mu$, where $\mu = \mu_{\boldsymbol{d,k,s}}$ is the polynomial map described in (2.5) and ℓ is a smooth function in matrix space (so $\mathcal{L}$ only depends on the end-to-end convolutional matrix $\overline{W} = \mu(\theta)$). This includes not only the square loss discussed further in section 6 but also classification losses (such as cross-entropy) or any other smooth function that depends only on the output of the LCN. A tuple of filters $\theta = (w_1, \ldots, w_L)$ is said to be *critical* for the loss $\mathcal{L}$ if all partial derivatives of $\mathcal{L}$ vanish at θ.

In subsection 5.1, we analyze critical points of the loss function, explaining their relation to polynomials with repeated roots. In subsection 5.2, we discuss the training dynamics and invariants in parameter space for LCNs, which allow us to formulate local Riemannian gradient dynamics in function space.

5.1. Critical points in parameter and function space. In this section, we discuss how critical points of an LCN correspond to polynomials that have repeated roots. We provide the proofs of the following proposition and theorems at the end of this section.

Proposition 5.1. *Let $\mathcal{L} = \ell \circ \mu$ be a loss function for an LCN with stride one, where ℓ is a smooth and convex function on the space of convolutional matrices. If $\theta = (w_1, \ldots, w_L)$ is critical for $\mathcal{L}$, then one of the following holds:*

1. *θ is a* global minimum *for $\mathcal{L}$, or*
2. *for some $i \neq j$, the homogeneous polynomials $\pi(w_i)$ and $\pi(w_j)$ have a common factor.*

Polynomials with multiple roots form in fact the boundaries of nonfilling architectures (as shown in Proposition 4.4); however, we will see that they arise as critical points more generally: *Every* critical point θ for the loss of an LCN corresponds to a critical point $\mu(\theta)$ of ℓ restricted to one of the multiple root loci Δ_λ, i.e., $\mu(\theta) \in \mathrm{Crit}(\ell|_{\pi^{-1}(\Delta_\lambda)})$, even if $\mu(\theta)$ does not necessarily lie on the boundary of the function space. In other words, $\mu(\theta)$ is a smooth point of $\pi^{-1}(\Delta_\lambda)$ and the differential $d\ell$ at $\mu(\theta)$ vanishes on the tangent space $T_{\mu(\theta)}\pi^{-1}(\Delta_\lambda)$.

Theorem 5.2. *Let $\mathcal{L} = \ell \circ \mu$ be a loss function for an LCN with stride one. Let $\theta = (w_1, \ldots, w_L) \in \mathrm{Crit}(\mathcal{L})$, and let $\overline{w} = \mu(\theta) \in \mathbb{R}^k$ be the end-to-end filter. Let $\lambda = (\lambda_1, \ldots, \lambda_r)$ be the partition of $k-1$ corresponding to the root structure of $\pi(\overline{w})$. Then $\overline{w} \in \mathrm{Crit}(\ell|_{\pi^{-1}(\Delta_\lambda)})$.*

To reverse Theorem 5.2, we characterize which critical points on the multiple root loci come from critical points of a given LCN.

Theorem 5.3. *Let $\mathcal{L} = \ell \circ \mu$ be a loss function for an LCN with stride one. Let λ be a partition of $k - 1$, and let $\overline{w} \in \mathrm{Crit}(\ell|_{\pi^{-1}(\Delta_\lambda)})$. If there exists $\theta = (w_1, \ldots, w_L)$ such that $\mu(\theta) = \overline{w}$ and each polynomial $\pi(w_1), \ldots, \pi(w_L)$ has no repeated roots, then $\theta \in \mathrm{Crit}(\mathcal{L})$.*

Moreover, if the loss ℓ in function space is sufficiently generic (e.g., if the training data is sufficiently generic), then we do not expect there to be any further critical points of the LCN other than the ones described in Theorem 5.3. We provide a formal argument for this assertion when $\ell = \ell_{X,Y}$ is the square loss (defined in (6.1)) with generic training data.

Theorem 5.4. *Let $X \in \mathbb{R}^{d_0 \times N}$ and $Y \in \mathbb{R}^{d_L \times N}$ be generic data matrices with $N \geq d_0$, and let $\mathcal{L} = \ell_{X,Y} \circ \mu$ be the square loss for an LCN with stride one. Then every $\theta = (w_1, \ldots, w_L) \in \mathrm{Crit}(\mathcal{L})$ satisfies that each polynomial $\pi(w_1), \ldots, \pi(w_L)$ has no repeated roots.*

To summarize, a critical point $\overline{w}$ of a generic loss ℓ on a multiple root locus $\pi^{-1}(\Delta_\lambda)$ comes from a critical point of the LCN if and only if $\pi(\overline{w})$ can be factored according to the LCN architecture such that no factor has a double root. Whether a polynomial with rrmp $(\rho \mid \gamma)$ can be factored according to the LCN architecture without double roots in any of the factors depends on a purely combinatorial property of $(\rho \mid \gamma)$ and the partition $(k_1 - 1, \ldots, k_L - 1)$ of $k-1$ that is given by the architecture. Indeed, the factorization of the polynomial is equivalent to the placing of ρ_i balls of size 1 and color i (for $1 \le i \le r$) and γ_j balls of size 2 and color $r + j$ (for $1 \le j \le c$) into L bins of sizes $k_1 - 1, \ldots, k_L - 1$ such that no bin contains two balls of the same color. If this placement of balls into bins is possible, we say that $(\rho \mid \gamma)$ is *compatible* with the LCN architecture.

All in all, the previous Theorems 5.2 to 5.4 show that the critical points of an LCN with a generic loss ℓ correspond exactly to the critical points on multiple root loci with an rrmp that is compatible with the architecture.

Corollary 5.5. *Let $X \in \mathbb{R}^{d_0 \times N}$ and $Y \in \mathbb{R}^{d_L \times N}$ be generic data matrices with $N \ge d_0$, and let $\mathcal{L} = \ell_{X,Y} \circ \mu$ be the square loss for an LCN with stride one. Let $(\rho \mid \gamma)$ be such that $\lambda = \lambda_{\rho|\gamma}$ is a partition of $k - 1$. If $(\rho \mid \gamma)$ is compatible with the LCN architecture, then for every $\overline{w} \in \mathrm{Crit}(\ell|_{\pi^{-1}(\Delta_\lambda)})$ with rrmp $(\rho \mid \gamma)$ there exists some $\theta \in \mu^{-1}(\overline{w}) \cap \mathrm{Crit}(\mathcal{L})$. Otherwise, every $\overline{w} \in \mathrm{Crit}(\ell|_{\pi^{-1}(\Delta_\lambda)})$ with rrmp $(\rho \mid \gamma)$ satisfies $\mu^{-1}(\overline{w}) \cap \mathrm{Crit}(\mathcal{L}) = \emptyset$.*

Example 5.6. We compare the LCN architectures with $\boldsymbol{k} = (3, 2, 2)$ and $\boldsymbol{k}' = (4, 2)$. We discussed in Example 4.3 that the function space associated with these architectures is the same. The polynomial $\pi(\overline{w}) = p_1 p_2^3$, where $p_i \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_1$, has rrmp $(13 \mid 0)$. We can find a factorization compatible with the architecture $\boldsymbol{k} = (3, 2, 2)$ where each polynomial associated with a filter does not have any repeated roots: $\pi(\overline{w}) = (p_1 p_2) \cdot p_2 \cdot p_2$. In contrast, for $\boldsymbol{k}' = (4, 2)$, any factorization of $\pi(\overline{w})$ according to the architecture will be such that the cubic polynomial associated with the filter of size 4 will have at least one pair of repeated roots.

For both architectures, we now list all compatible $(\rho \mid \gamma)$ by providing a feasible factorization; the remaining ones are marked with a dash:

$\rho\|\gamma$	$1111\|0$	$112\|0$	$22\|0$	$13\|0$	$4\|0$	$11\|1$	$2\|1$	$0\|2$	$0\|11$
$(3, 2, 2)$	$p_1 p_2 \cdot p_3 \cdot p_4$	$p_1 p_2 \cdot p_3 \cdot p_3$	$p_1 p_2 \cdot p_1 \cdot p_2$	$p_1 p_2 \cdot p_2 \cdot p_2$	$-$	$q_1 \cdot p_1 \cdot p_2$	$q_1 \cdot p_1 \cdot p_1$	$-$	$-$
$(4, 2)$	$p_1 p_2 p_3 \cdot p_4$	$p_1 p_2 p_3 \cdot p_3$	$-$	$-$	$-$	$p_1 q_1 \cdot p_2$	$p_1 q_1 \cdot p_1$	$-$	$-$

To prove the theorems above, we make use of the composition rule of differentials:

$$d\mathcal{L} = d\ell \circ d\mu_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}}.$$

Even if the parameterization map μ is surjective (i.e., the function space $\mathcal{M} = \mathcal{M}_{\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s}}$ is filling), the differential $d\mu$ (equivalently, the Jacobian matrix of μ) might not have maximal rank everywhere. This implies that the parameterization may give rise to critical points for $\mathcal{L}$ in addition to the global minimum even if the function space $\mathcal{M}$ is a vector space and ℓ is convex. Following [35], we call this kind of critical point *spurious*. In contrast, a critical point θ for $\mathcal{L}$ is *pure* if $\mu(\theta)$ is a critical point of the loss in function space, i.e., of $\ell|_{\mathcal{M}}$.

Given an end-to-end filter $\overline{w} \in \mathbb{R}^k$ of an LCN, we describe the set of parameters θ such that $\mu(\theta) = \overline{w}$, i.e., the fiber of the parameterization. We say that two parameters $\theta = (w_1, \ldots, w_L)$ and $\theta' = (w_1', \ldots, w_L')$ are equivalent up to scalar factors if there exists $(\kappa_1, \ldots, \kappa_L) \in (\mathbb{R}^*)^L$ with $\prod_{i=1}^L \kappa_i = 1$ such that $w_i = \kappa_i w_i'$.

Lemma 5.7. *Let $\overline{w} \in \mathbb{R}^k \setminus \{0\}$ be an end-to-end filter of an LCN with stride one. Then the fiber $\mu^{-1}(\overline{w})$ consists of a finite set of scaling equivalence classes. Moreover, either all points in the same equivalence class are critical or none are.*

Proof. The fact that $\mu^{-1}(\overline{w})$ consists of a finite set of scaling equivalence classes follows from the uniqueness of the decomposition of $\pi(\overline{w}) \in \mathbb{R}[\mathsf{x}, \mathsf{y}]$ into irreducible factors: Each equivalence class corresponds to a way of aggregating factors of $\pi(\overline{w})$ into polynomials of degrees $k_1 - 1, \ldots, k_L - 1$, where $(k_1, \ldots, k_L)$ are the filter widths of the LCN. For the second claim, we use the fact that if θ and θ' are equivalent under rescaling, then the image of the differential $d\mu$ at these points is the same. This will follow from our characterization of the image of $d\mu$ in (5.1). As a result, $d\mathcal{L}(\theta) = d\ell(\overline{w}) \circ d\mu(\theta) = 0$ if and only if $d\mathcal{L}(\theta') = d\ell(\overline{w}) \circ d\mu(\theta') = 0$. ∎

Example 5.8. We discuss the parameters corresponding to particular points in the function space of the LCN with $\boldsymbol{k} = (2, 2)$ from Example 2.4. Using the identification π with polynomials, a point in the interior of the function space is a quadratic polynomial with two distinct real roots: $P = p_1 p_2$, where $p_1 = a\mathsf{x} + b\mathsf{y}$ and $p_2 = c\mathsf{x} + d\mathsf{y}$ are not scalar multiples of each other. The fiber $\mu^{-1}(P)$ of the parameterization has four connected one-dimensional components: $\{(\kappa p_1, \frac{1}{\kappa}p_2) : \kappa > 0\}$, $\{(\kappa p_1, \frac{1}{\kappa}p_2) : \kappa < 0\}$, $\{(\kappa p_2, \frac{1}{\kappa}p_1) : \kappa > 0\}$, and $\{(\kappa p_2, \frac{1}{\kappa}p_1) : \kappa < 0\}$.

The points on the boundary of the function space are exactly those $P = p_1 p_2$, where p_1 and p_2 are linearly dependent. In other words, the parameters mapping to boundary points are those which satisfy $bc - ad = 0$. We will see in Proposition 5.11 that these are exactly the parameters where the Jacobian of the parameterization μ drops rank.

Hence, a nonzero point on the boundary of the function space is of the form $P = \pm p^2$ for some linear term $p \neq 0$. Thus, the fiber $\mu^{-1}(P)$ has two connected one-dimensional components: $\{(\pm \kappa p, \frac{1}{\kappa}p) : \kappa > 0\}$ and $\{(\pm \kappa p, \frac{1}{\kappa}p) : \kappa < 0\}$. Finally, the fiber $\mu^{-1}(0)$ over the origin is the union of the two planes $\{(p, 0) : p \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_1\}$ and $\{(0, p) : p \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_1\}$.

Remark 5.9. In the case of fully connected linear networks, spurious critical points are always *saddles*. More precisely, for a fully connected network, a loss function $\mathcal{L} = \ell \circ \mu$ where ℓ is a smooth convex function has nonglobal minima if and only if the restriction of ℓ on the function space has nonglobal minima [35, Proposition 10]. In Example 5.10, we will see that this is *not* true for LCNs, and it is possible for the loss function $\mathcal{L}$ to have nonglobal minima even if the function space is filling.

The intuitive geometric reason for this different behavior between fully connected and convolutional linear networks lies in the very different structure of the fibers. For fully connected networks, as shown in [35, Proposition 9], it is possible to perturb a spurious critical point within the fiber to obtain another point that is not a critical point for the loss. This in turn implies that the original spurious critical point could not have been a minimum for the loss. In the case of LCNs, however, this argument does not apply by Lemma 5.7. Thus, for LCNs, it is possible for spurious critical points to be (nonglobal) local minima.

Example 5.10. Let us consider an LCN with filter sizes $\boldsymbol{k} = (2, 3)$. This is a filling architecture. In terms of polynomials, its function space is the set of all cubic polynomials. The two layers of the LCN factor such a polynomial into a linear and a quadratic term. We fix the target polynomial $u = (\mathsf{x} + \mathsf{y})(\mathsf{x}^2 + 1/10\,\mathsf{y}^2)$. In the space of polynomials $\mathbb{R}[\mathsf{x}, \mathsf{y}]_3$, we use the loss $\ell(p)$ that is the squared sum of the coefficients of $p - u$. We will see that $\mathcal{L} = \ell \circ \mu$

has a nonglobal minimum, although ℓ clearly only has one local and one global minimum.

Almost every cubic polynomial p can be factored as $p = q_1 q_2$, where $q_1 = \mathsf{x} + a\mathsf{y}$ and $q_2 = b\mathsf{x}^2 + c\mathsf{xy} + d\mathsf{y}^2$. We assume that q_1 is monic (i.e., the coefficient for x is 1) since there is otherwise a scaling ambiguity among q_1 and q_2 which does not affect the product $q_1 q_2$ or the property of critical points of $\mathcal{L}$. Our objective function $\mathcal{L}$ thus becomes

$$\mathcal{L} : \mathbb{R}^4 \longrightarrow \mathbb{R}; \quad (a,b,c,d) \longmapsto (b-1)^2 + (ab+c-1)^2 + (ac+d-\text{\textonesuperior/}10)^2 + (ad-\text{\textonesuperior/}10)^2.$$

A direct computation reveals that $\mathcal{L}$ has 10 critical points over the complex numbers: Three of them (1 real, 2 complex) correspond to factorizations of the target polynomial u, so that $\mathcal{L} = 0$. The other seven (3 real, 4 complex) are spurious critical points and yield factorizations of seven different cubic polynomials. One of the latter three real spurious critical points is

$$a \approx 0.0578445483987, b \approx 1.0000187825172, c \approx 0.941829719725, d \approx 0.0511336556138.$$

The Hessian of $\mathcal{L}$ at this point is positive definite, showing that it is indeed a local minimum.

We now investigate the existence of spurious critical points for LCNs by studying the rank of the differential of the parameterization map μ.

Proposition 5.11. *Let μ be the parameterization map of an LCN with stride one. Given filter weights $\theta = (w_1, \ldots, w_L)$, we write $\overline{w} = \mu(\theta) \in \mathbb{R}^k$ for the end-to-end filter. Using the identification π, the image of the differential $d\mu$ at the point θ is the set of polynomials*

$$(5.1) \qquad\qquad H(\theta) := \big\{ h\,g : h \in \mathbb{R}[\mathsf{x},\mathsf{y}]_{k-1-\deg(g)} \big\} \subset \mathbb{R}[\mathsf{x},\mathsf{y}]_{k-1},$$

where $g = \gcd\big(\frac{\pi(\overline{w})}{\pi(w_1)}, \ldots, \frac{\pi(\overline{w})}{\pi(w_L)}\big) \in \mathbb{R}[\mathsf{x},\mathsf{y}]$. In particular, $d\mu(\theta)$ is surjective if and only if $\gcd(\pi(w_i), \pi(w_j)) = 1$ for every pair $i, j \in \{1, \ldots, L\}$ with $i \neq j$.

To prove this result, we make use of the following homogeneous version of Bezout's identity.

Lemma 5.12. *Let $p_1, \ldots, p_n$ be homogeneous polynomials in $\mathbb{R}[\mathsf{x},\mathsf{y}]$ of arbitrary degree, and let $g = \gcd(p_1, \ldots, p_n)$ and $l = \mathrm{lcm}(p_1, \ldots, p_n)$. For any $d \geq \deg(l) - 1$, the following sets of homogeneous polynomials in $\mathbb{R}[\mathsf{x},\mathsf{y}]_d$ are equal:*

$$I_d := \{\alpha_1 p_1 + \cdots + \alpha_n p_n : \alpha_i \in \mathbb{R}[\mathsf{x},\mathsf{y}]_{d-\deg(p_i)}\}, \quad J_d := \{hg : h \in \mathbb{R}[\mathsf{x},\mathsf{y}]_{d-\deg(g)}\}.$$

Proof. If $g = l$ (i.e., $p_1 = \cdots = p_n$), the assertion is clear. Hence, we assume that $\deg(g) < \deg(l)$. We always have that $I_d \subset J_d$ since any algebraic combination of $p_1, \ldots, p_n$ is a multiple of g. For the converse, we observe that it is enough to prove that $J_{\deg(l)-1} \subset I_{\deg(l)-1}$ (if hg belongs to the ideal generated by $p_1, \ldots, p_n$, then so do $\mathsf{x}hg$ and $\mathsf{y}hg$). Moreover, it is sufficient to assume that the linear form y is not a factor of l (otherwise, we apply a general invertible linear transformation to the input variables, e.g., $\mathsf{x}' = \mathsf{x}, \mathsf{y}' = \mathsf{y} + t\mathsf{x}$ for appropriate t). We can thus safely *dehomogenize* and obtain univariate polynomials $\tilde{p}_i = p_i(\mathsf{x},1)$, $\tilde{g} = g(\mathsf{x},1)$, $\tilde{l} = l(\mathsf{x},1)$ in $\mathbb{R}[\mathsf{x}]$. Note that these polynomials have the same degrees as their homogeneous counterparts and also that $\tilde{g} = \gcd(\tilde{p}_1, \ldots, \tilde{p}_n)$ and $\tilde{l} = \mathrm{lcm}(\tilde{p}_1, \ldots, \tilde{p}_n)$. By Bezout's identity, we can write $\tilde{\beta}_1 \tilde{p}_1 + \cdots + \tilde{\beta}_n \tilde{p}_n = \tilde{g}$ for some $\tilde{\beta}_i \in \mathbb{R}[\mathsf{x}]$. Thus, for any $\tilde{h} \in \mathbb{R}[\mathsf{x}]$, we get a decomposition $\tilde{\alpha}_1 \tilde{p}_1 + \cdots + \tilde{\alpha}_n \tilde{p}_n = \tilde{h}\tilde{g}$. We now claim that we can assume

$\deg(\tilde{\alpha}_i) < \deg(\tilde{l}) - \deg(\tilde{p}_i)$ whenever $\deg(\tilde{h}\tilde{g}) \leq \deg(\tilde{l}) - 1$. Indeed, if this is not the case, then we write $\tilde{\alpha}_i = \tilde{q}_i \tilde{p}_i^* + \tilde{r}_i$ with $\tilde{p}_i^* = \tilde{l}/\tilde{p}_i$ and $\deg(\tilde{r}_i) < \deg(\tilde{p}_i^*) = \deg(\tilde{l}) - \deg(\tilde{p}_i)$, and we obtain $(\sum_{i=1}^n \tilde{q}_i)\tilde{l} + \tilde{r}_1\tilde{p}_1 + \cdots + r_n\tilde{p}_n = \tilde{h}\tilde{g}$. Since $\deg(\tilde{h}\tilde{g}) < \deg(\tilde{l})$, we deduce that $\sum_{i=1}^n \tilde{q}_i = 0$, which yields a decomposition with the desired degree bounds. We can now *homogenize* this relation so that the total degree is $\deg(l) - 1 = \deg(\tilde{l}) - 1$ to obtain $\alpha_1 p_1 + \cdots + \alpha_n p_n = hg$, where

$$\alpha_i(\mathsf{x},\mathsf{y}) = \mathsf{y}^{\deg(\tilde{l})-1-\deg(\tilde{p}_i)}\tilde{\alpha}_i\left(\frac{\mathsf{x}}{\mathsf{y}}\right), \quad h(\mathsf{x},\mathsf{y}) = \mathsf{y}^{\deg(\tilde{l})-1-\deg(\tilde{g})}\tilde{h}\left(\frac{\mathsf{x}}{\mathsf{y}}\right),$$

$$p_i(\mathsf{x},\mathsf{y}) = \mathsf{y}^{\deg(\tilde{p}_i)}\tilde{p}_i\left(\frac{\mathsf{x}}{\mathsf{y}}\right), \qquad\qquad g(\mathsf{x},\mathsf{y}) = \mathsf{y}^{\deg(\tilde{g})}\tilde{g}\left(\frac{\mathsf{x}}{\mathsf{y}}\right).$$

Note that p_i and g are the original homogeneous polynomials from the statement. Since $\tilde{h}$ can be chosen arbitrarily in $\mathbb{R}[\mathsf{x}]_{\leq \deg(\tilde{l})-1-\deg(\tilde{g})}$, we have that $h(\mathsf{x},\mathsf{y})$ can be an arbitrary polynomial in $\mathbb{R}[\mathsf{x},\mathsf{y}]_{\deg(l)-1-\deg(g)}$. We thus conclude that $J_{\deg(l)-1} \subset I_{\deg(l)-1}$. ∎

Proof of Proposition 5.11. For notational simplicity, we omit writing the map π and identify $\mathbb{R}^k$ with $\mathbb{R}[\mathsf{x},\mathsf{y}]_{k-1}$. The differential of the parameterization can be written as

$$(5.2) \qquad d\mu(\theta)\colon \mathbb{R}^{k_1+\cdots+k_L} \longrightarrow \mathbb{R}^k; \quad (\dot{w}_1,\ldots,\dot{w}_L) \longmapsto \dot{w}_1\frac{\overline{w}}{w_1} + \cdots + \dot{w}_L\frac{\overline{w}}{w_L},$$

where the product means polynomial multiplication and each $\dot{w}_i$ is simply a symbol for the coordinates of the tangent vector corresponding to the variable w_i. Since $\overline{w}$ is a common multiple of $\frac{\overline{w}}{w_1},\ldots,\frac{\overline{w}}{w_L}$, using Lemma 5.12 we deduce that the image of (5.2) is equal to $H(\theta)$ in (5.1). This shows that the differential is surjective if and only if $\gcd\left(\frac{\overline{w}}{w_1},\ldots,\frac{\overline{w}}{w_L}\right) = 1$. The latter is equivalent to $\gcd(w_i, w_j) = 1$ for every pair $i \neq j$. Indeed, any common factor of w_i, w_j is a factor of each $\frac{\overline{w}}{w_m}$; conversely, if q is an irreducible common factor of $\frac{\overline{w}}{w_1},\ldots,\frac{\overline{w}}{w_L}$, then $q|w_i$ for some $i \in \{1,\ldots,L\}$, and from $q|\frac{\overline{w}}{w_i}$ we deduce $q|w_j$ for some $j \neq i$. ∎

We are finally ready to prove the main theorems of this section. As in the proof of Proposition 5.11, we omit writing the map π and identify filters with polynomials.

Proof of Proposition 5.1. Let $\overline{W} = \mu(w_1,\ldots,w_L)$. If $(d\ell \circ d\mu)(\theta) = 0$, then either $d\ell(\overline{W}) = 0$ or $d\mu(\theta)$ must not have full rank. The assumption that ℓ is convex means that if $d\ell(\overline{W}) = 0$, then $\overline{W}$ is a global minimum for ℓ and thus θ is a global minimum for $\mathcal{L}$. Alternatively, if $d\mu(\theta)$ does not have full rank, then by Proposition 5.11 the second condition holds. ∎

Proof of Theorem 5.2. If $\lambda = (1,\ldots,1)$, then $\overline{w}$ has $k-1$ distinct roots. By Proposition 5.11, the differential $d\mu(\theta)$ has full rank; i.e., θ is a pure critical point. This in turn implies that $\overline{w}$ is a critical point for ℓ in $\Delta_\lambda = \mathbb{R}[x,y]_{k-1}$.

We next assume that $\lambda = (\lambda_1,\ldots,\lambda_r)$ is such that $\lambda_1 > 1$. Let $\overline{w} = q_1^{\lambda_1}\cdots q_r^{\lambda_r}$ be the decomposition of $\overline{w}$ into pairwise distinct complex linear forms $q_i \in \mathbb{C}[\mathsf{x},\mathsf{y}]_1$. Note that if $\overline{w} = \prod_{i=1}^r q_i^{\lambda_i}$ is real, then so is $\prod_{i=1}^r q_i^{\lambda_i-1}$. We claim that the (real) tangent space of Δ_λ at the smooth point $\overline{w}$ is given by

$$(5.3) \qquad T_{\overline{w}}\Delta_\lambda = \left\{ h\prod_{i=1}^r q_i^{\lambda_i-1} \ : \ h \in \mathbb{R}[\mathsf{x},\mathsf{y}]_r \right\} \subset \mathbb{R}[\mathsf{x},\mathsf{y}]_{k-1}.$$

Indeed, let us assume without loss of generality that $q_{i_1}, \ldots, q_{i_{r'}}$ are real linear forms, while $q_{j_1}, q_{j_1+1}, \ldots, q_{j_{r''}}, q_{j_{r''}+1}$ are such that each (q_{j_l}, q_{j_l+1}) is a conjugate pair of complex linear forms, with $r' + 2r'' = r$ (in particular, we must also have $\lambda_{j_l} = \lambda_{j_l+1}$ for all $l = 1, \ldots, r''$). We write $t_{j_l} = q_{j_l} q_{j_l+1}$ for the irreducible real quadratic form that is the product of the pair q_{j_l}, q_{j_l+1}. In a neighborhood of $\overline{w}$, the multiple root locus Δ_λ is the image of

$$(5.4) \qquad (q_{i_1}, \ldots, q_{i_{r'}}, t_{j_1}, \ldots, t_{j_{r''}}) \mapsto \prod_{m=1}^{r'} q_{i_m}^{\lambda_{i_m}} \prod_{l=1}^{r''} t_{j_l}^{\lambda_{j_l}}$$

as $q_{i_m} \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_1$ and $t_{j_l} \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_2$ vary. Arguing as in the proof of Proposition 5.11 (based on Lemma 5.12), we have that the image of the differential of (5.4) is the set of homogeneous polynomials of degree $k - 1$ that are multiples of $g' = \gcd\left(\frac{\overline{w}}{q_{i_1}}, \ldots, \frac{\overline{w}}{q_{i_{r'}}}, \frac{\overline{w}}{t_{j_1}}, \ldots, \frac{\overline{w}}{t_{j_{r''}}}\right)$. We finally observe that $g' = \prod_{i=1}^r q_i^{\lambda_i - 1}$.

Having shown (5.3), we recall that the image of the differential of the parameterization μ is $\mathrm{Im}(d\mu(\theta)) = \left\{hg : h \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1-\deg(g)}\right\} \subset \mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1}$, where g is as in Proposition 5.11. To conclude the proof, we note that $T_{\overline{w}}\Delta_\lambda \subset \mathrm{Im}(d\mu(\theta))$ since $g \mid \prod_{i=1}^r q_i^{\lambda_i - 1}$. Indeed, the linear form q_i can appear in the factorization of g with multiplicity at most $\lambda_i - 1$, for if $q_i | w_j$, then $q_i^{\lambda_i} \nmid \frac{\overline{w}}{w_j}$. The fact that $T_{\overline{w}}\Delta_\lambda \subset \mathrm{Im}(d\mu(\theta))$ implies the statement of the theorem since $0 = d\mathcal{L}(\theta) = d\ell(\overline{w}) \circ d\mu(\theta)$ implies $d\ell(\overline{w})|_{T_{\overline{w}}\Delta_\lambda} = 0$, i.e., $\overline{w}$ is a critical point of $\ell|_{\Delta_\lambda}$. ∎

Proof of Theorem 5.3. Below we show the following claim for $\overline{w}$ being a smooth point of Δ_λ: For every $\theta \in \mu^{-1}(\overline{w})$, it holds that $\mathrm{Im}(d\mu(\theta)) = T_w(\Delta_\lambda)$ if and only if each w_i has no repeated roots. Hence, under the assumptions in the theorem, $d\mathcal{L}(\theta) = d\ell(\overline{w}) \circ d\mu(\theta) = d\ell(\overline{w})|_{T_{\overline{w}}\Delta_\lambda} = 0$.

To prove our claim, we use the fact that $T_{\overline{w}}(\Delta_\lambda) = \{h \prod_{i=1}^r q_i^{\lambda_i - 1} : h \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_r\}$ as in (5.3) if $\overline{w} = \prod_{i=1}^r q_i^{\lambda_i}$ and that $\mathrm{Im}(d\mu(\theta)) = \left\{h\,g : h \in \mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1-\deg(g)}\right\}$, where g is the gcd defined in Proposition 5.11. We observe that w_i has no repeated roots if and only if $\prod_{i=1}^r q_i^{\lambda_i - 1}|(\overline{w}/w_i)$. This happens for every $i = 1, \ldots, L$ if and only if $\prod_{i=1}^r q_i^{\lambda_i - 1}|g$. Since $g| \prod_{i=1}^r q_i^{\lambda_i - 1}$ always holds, as seen in the proof of Theorem 5.2, the latter is equivalent to $\mathrm{Im}(d\mu(\theta)) = T_{\overline{w}}(\Delta_\lambda)$. ∎

Proof of Theorem 5.4. As we will see in (6.1)–(6.3), if $XX^\top$ has full rank, any square loss $\ell_{X,Y}(\overline{w})$ is equivalent up to a constant additive term to a function of the form $\ell_{\Sigma,u}(\overline{w}) = \|u - \overline{w}\|_\Sigma^2$, where u is a target filter and $\langle \cdot, \cdot \rangle_\Sigma$ is a nondegenerate inner product on $\mathbb{R}^k$. Here k is the size of the filters u and $\overline{w}$. We now show that for *arbitrary* inner products $\langle \cdot, \cdot \rangle_\Sigma$ and *generic* target filters u any critical point $\theta = (w_1, \ldots, w_L)$ of $\mathcal{L}_{\Sigma,u} = \ell_{\Sigma,u} \circ \mu$ is such that each polynomial $w_1, \ldots, w_L$ has no repeated roots.[3] Fixing the inner product arbitrarily, we consider for all $i = 1, \ldots, L$ the set

$$Z_i = \{(\theta, \overline{w}, u) : \theta \in \mathrm{Crit}(\mathcal{L}_{\Sigma,u}), \mu(\theta) = \overline{w}, \theta = (w_1, \ldots, w_L), w_i \text{ has a repeated root}\}.$$

---

[3]In the setting of subsection 6.1, we need to observe that the inner product only depends on the input data matrix $X \in \mathbb{R}^{d_0 \times N}$ and that as the output data matrix $Y \in \mathbb{R}^{d_L \times N}$ varies the target matrix $U = YX^\top(XX^\top)^{-1}$ can be an arbitrary matrix of size $d_L \times d_0$ due to the assumption $N \geq d_0$. In particular, for any nondegenerate inner product $\langle \cdot, \cdot \rangle_\Sigma$, all possible target filters u arise.

Note that this is an algebraic set. We claim that $\dim(\mathrm{pr}_3(Z_i)) \leq k-1$ holds for all $i = 1, \ldots, L$, where pr_3 denotes the projection on the third factor. This implies the statement of the theorem since it shows that the set of target filters u such that any of the corresponding critical points has a filter with repeated roots has positive codimension in $\mathbb{R}^k$. To prove our claim on the dimension of Z_i, we further consider the sets

$$Z_{i,\lambda} = \{(\theta, \overline{w}, u) : (\theta, \overline{w}, u) \in Z_i, \ \overline{w} \in \mathrm{Reg}(\Delta_\lambda)\},$$

where λ is a partition of $k-1$ and $\mathrm{Reg}(\Delta_\lambda)$ denotes the locus of smooth points on Δ_λ. Clearly, $Z_i = \bigcup_\lambda Z_{i,\lambda}$. As shown in the proof of Theorem 5.3, if $\overline{w}$ is a smooth point of Δ_λ and w_i has a repeated root, then $T_{\overline{w}}\Delta_\lambda \subsetneq \mathrm{Im}(d\mu(\theta))$. The fiber $\mu^{-1}(\overline{w})$ consists of finitely many scaling equivalence classes (see Lemma 5.7), and all θ in the same equivalence class yield the same $\mathrm{Im}(d\mu(\theta))$. We note that $\theta \in \mathrm{Crit}(\mathcal{L}_{\Sigma,u})$ means equivalently that $u - \overline{w} \perp_\Sigma \mathrm{Im}(d\mu(\theta))$, or in other words $u \in \overline{w} + \mathrm{Im}(d\mu(\theta))^{\perp_\Sigma}$. We write $N_{\Sigma,\overline{w}}$ for the finite union of the affine spaces $\overline{w} + \mathrm{Im}(d\mu(\theta))^{\perp_\Sigma}$ where $\theta = (w_1, \ldots, w_L) \in \mu^{-1}(\overline{w})$ and some w_i has a repeated root. Hence, $u \in \mathrm{pr}_3(Z_{i,\lambda})$ if and only if there is some smooth point $\overline{w}$ on Δ_λ such that $u \in N_{\Sigma,\overline{w}}$. Since $\dim(N_{\Sigma,\overline{w}}) < \dim((T_{\overline{w}}\Delta_\lambda)^{\perp_\Sigma}) = k - \dim(\Delta_\lambda)$, we see that $\dim(\mathrm{pr}_3(Z_{i,\lambda})) \leq \dim(\Delta_\lambda) + \max_{\overline{w} \in \mathrm{Reg}(\Delta_\lambda)} \dim(N_{\Sigma,\overline{w}}) < k$. This concludes the proof. ∎

5.2. Training dynamics. In this section, we discuss some matters related to the training dynamics of LCNs. We begin by proving the existence of dynamical invariants in parameter space. Similar results have been shown for linear networks [6] and ReLU networks [36]. We restrict ourselves to networks with stride one.

Proposition 5.13. Let $\mathcal{L} = \ell \circ \mu$ be a loss function for an LCN with stride one, where ℓ is a smooth function on the space of convolutional matrices. Let $\theta(t)$ be an integral curve in parameter space for the negative gradient field of $\mathcal{L}$, i.e., $\dot{\theta}(t) = -\nabla(\mathcal{L}(\theta(t)))$. If we write $\theta(t) = (w_1(t), \ldots, w_L(t))$, where w_i are filter vectors, then the quantities $\delta_{ij}(t) = \|w_i(t)\|^2 - \|w_j(t)\|^2$ for $i, j \in \{1, \ldots, L\}$ remain constant for all $t \in \mathbb{R}$.

Proof. For any $\kappa \in \mathbb{R} \setminus \{0\}$, let $\nu_{ij}(\theta, \kappa) = (w_1, \ldots, \kappa w_i, \ldots, {}^1/\kappa\, w_j, \ldots, w_L)$. Because the map μ is multilinear in the filters, we have that $\mathcal{L}(\nu_{ij}(\theta, \kappa))$ does not depend on κ. This implies that $\frac{d}{d\kappa}\mathcal{L}(\nu_{ij}(\theta, \kappa)) \equiv 0$ and in particular

$$0 = \left[\frac{d}{d\kappa}\mathcal{L}(\nu_{ij}(\theta, \kappa))\right]_{\kappa=1} = \frac{d}{d\theta}\mathcal{L}(\theta) \cdot \left[\frac{d}{d\kappa}\nu_{ij}(\theta, \kappa)\right]_{\kappa=1} = \frac{d}{d\theta}\mathcal{L}(\theta) \cdot [0, \ldots, w_i, \ldots, -w_j, \ldots 0]^\top.$$

We now observe that

$$\frac{d}{dt}\delta_{ij}(t) = \frac{d}{d\theta}\delta_{ij} \cdot \frac{d}{dt}\theta(t) = -2\,[0, \ldots, w_i, \ldots, -w_j, \ldots, 0] \cdot \frac{d}{d\theta}\mathcal{L}(\theta) = 0.$$

This shows that δ_{ij} remains constant for all t. ∎

As discussed in Lemma 5.7, each connected component of the fiber $\mu^{-1}(p)$ of a polynomial p under the parameterization map μ consists of the different rescalings of a fixed factorization of p. Proposition 5.13 can be used to eliminate the scaling ambiguity, assuming that the parameters (or the invariants) are known at initialization.

Corollary 5.14. *Let μ be the parameterization map of an LCN with stride one. Given filters $\theta = (w_1, \ldots, w_L)$, the set of parameter values θ' such that $\mu(\theta) = \mu(\theta')$ and $\delta_{i,i+1} = \delta'_{i,i+1}$ for all $i = 1, \ldots, L - 1$ is finite, where δ_{ij} and δ'_{ij} are the invariants associated with θ and θ', respectively, as defined in Proposition 5.13.*

Proof. Let $q_i = \pi(w_i)$ be the polynomials associated with the filters w_i, and let $p = q_L \cdots q_1 = \pi(\mu(\theta))$. As in Lemma 5.7, the factors q_i are determined by p up to a finite number of root permutations and also up to rescalings: $q'_i = \kappa_i q_i$, where $\prod_{i=1}^{L} \kappa_i = 1$. It remains to show that there are only finitely many choices of the scale factors κ_i such that $\delta_{i+1,i} = \|\kappa_{i+1} q_{i+1}\|^2 - \|\kappa_i q_i\|^2$ for $i = 1, \ldots, L - 1$. We can recover the possible κ_i by solving a simple polynomial system. Setting $\beta_i = \|\kappa_i q_i\|^2 = \kappa_i^2 \|q_i\|^2$ and $\delta_i = \delta_{i+1,i}$, we have that

$$\beta_{i+1} - \beta_i = \delta_i, \ (i = 1, \ldots, L - 1), \quad \prod_{i=1}^{L} \beta_i = H,$$

where $H = \prod_{i=1}^{L} \|q_i\|^2$. In particular, all β_i are determined by β_1: $\beta_{i+1} = \beta_1 + \delta_1 + \cdots + \delta_i$ for all $i = 1, \ldots, L - 1$. Hence, we obtain one polynomial equation of degree L in β_1, namely

$$(5.5) \qquad \beta_1 \cdot (\beta_1 + \delta_1) \cdots (\beta_1 + \delta_1 + \cdots + \delta_{L-1}) - H = 0.$$

There are finitely many real solutions for β_1 in (5.5) such that $\beta_1, \ldots, \beta_1 + \delta_1 + \cdots + \delta_{L-1}$ are all positive. Each such solution gives finitely many choices for the κ_i. ∎

The previous statement can be used to study the impact of the initialization (within the fiber) on the training dynamics. More precisely, we can compute the Riemannian metric that locally characterizes the gradient dynamics in function space. In general, if $\mathcal{L} = \ell \circ \mu$ and $\theta(t)$ is an integral curve for the negative gradient field of $\mathcal{L}$, then setting $\overline{W}(t) = \mu(\theta(t))$ we have

$$(5.6) \qquad \frac{d}{dt}\overline{W}(t) = -K(\theta(t)) \cdot \nabla\ell(\overline{W}(t)),$$

where $K(\theta) = \mathrm{Jac}_\mu(\theta) \mathrm{Jac}_\mu(\theta)^\top$ and $\mathrm{Jac}_\mu(\theta)$ is the Jacobian matrix of μ at θ. The matrix $K(\theta)$ corresponds to the *neural tangent kernel* [23] popular in the analysis of neural networks and can be viewed as a "preconditioning" matrix that modifies the gradient of ℓ in function space. If $K(\theta)$ could be written as a function of $\overline{W} = \mu(\theta)$, then we could view (5.6) as a Riemannian gradient flow for the metric tensor K^{-1} (this assumes K is not singular, which is generically the case; note that K is always positive semidefinite). In general, however, the matrix $K(\theta)$ depends on the parameter θ, and different parameters within the fiber of $\overline{W}$ will yield different matrices. In particular, the trajectories in function space are susceptible to the relative scalings of the filters at initialization. Using the dynamical invariants, however, we can bypass this problem by restricting ourselves to the parameters that are dynamically reachable, defining a local metric in function space associated with the initialization. We show this idea with a simple example.

Example 5.15. Let us consider the factorization of a cubic polynomial into a linear and a quadratic factor. We write $C = A \cdot B$, where

$$C = c_3 \mathsf{x}^3 + c_2 \mathsf{x}^2 \mathsf{y} + c_1 \mathsf{x} \mathsf{y}^2 + c_0 \mathsf{y}^3, \quad A = a_2 \mathsf{x}^2 + a_1 \mathsf{x} \mathsf{y} + a_0 \mathsf{y}^2, \quad B = b_1 \mathsf{x} + b_0 \mathsf{y}.$$

The Jacobian matrix of the parameterization map $\mu : (a_2, a_1, a_0, b_1, b_0) \mapsto (c_3, c_2, c_1, c_0) = (a_2 b_1, a_2 b_0 + a_1 b_1, a_1 b_0 + a_0 b_1, a_0 b_0)$ is

$$\mathrm{Jac}_\mu(a_2, a_1, a_0, b_1, b_0) = \begin{bmatrix} b_1 & 0 & 0 & a_2 & 0 \\ b_0 & b_1 & 0 & a_1 & a_2 \\ 0 & b_0 & b_1 & a_0 & a_1 \\ 0 & 0 & b_0 & 0 & a_0 \end{bmatrix} \in \mathbb{R}^{4 \times 5}.$$

The matrix $K(\theta) = \mathrm{Jac}_\mu \mathrm{Jac}_\mu^\top \in \mathbb{R}^{4 \times 4}$ can be written as $K(\theta) = K_B + K_A$, where

$$K_B = \begin{bmatrix} b_1^2 & b_0 b_1 & 0 & 0 \\ b_0 b_1 & b_0^2 + b_1^2 & b_0 b_1 & 0 \\ 0 & b_0 b_1 & b_0^2 + b_1^2 & b_0 b_1 \\ 0 & 0 & b_0 b_1 & b_0^2 \end{bmatrix}, \quad K_A = \begin{bmatrix} a_2^2 & a_1 a_2 & a_0 a_2 & 0 \\ a_1 a_2 & a_1^2 + a_2^2 & a_0 a_1 + a_1 a_2 & a_0 a_2 \\ a_0 a_2 & a_0 a_1 + a_1 a_2 & a_0^2 + a_1^2 & a_0 a_1 \\ 0 & a_0 a_2 & a_0 a_1 & a_0^2 \end{bmatrix}.$$

This matrix clearly depends on the coefficients of A and B; however, we can recover it given only C and the value of $\delta = \|A\|^2 - \|B\|^2$, which is fixed from the initialization. To resolve the scale ambiguity, we replace A, B with $\gamma A, 1/\gamma B$ with $\gamma \in \mathbb{R} \setminus \{0\}$ (here A and B can be fixed arbitrarily so that $A \cdot B = C$). We see that $K(\theta)$ becomes a one-dimensional (quadratic) family of matrices $K(\gamma) = (1/\gamma)^2 K_B + \gamma^2 K_A$. Using $\|\gamma A\|^2 - \|1/\gamma B\|^2 = \delta$ and writing $\beta = \|1/\gamma B\|^2$ and $H = \|A\|^2 \|B\|^2$, we deduce that

$$\beta^2 + \delta\beta - H = 0, \quad \beta > 0, \beta + \delta > 0, H > 0,$$

which is the same as (5.5). This yields $\beta = \frac{-\delta + \sqrt{\delta^2 + 4H}}{2}$ and $\gamma^2 = \frac{2\|B\|^2}{-\delta + \sqrt{\delta^2 + 4H}}$.

6. Optimization of the square loss.
In this section, we investigate the minimization of quadratic losses and the Euclidean distance degrees for LCNs.

6.1. The square loss.
Given data $X \in \mathbb{R}^{d_0 \times N}$ and $Y \in \mathbb{R}^{d_L \times N}$ (where N is the number of data samples), and assuming that $XX^\top$ has full rank, we write the quadratic loss as

$$\begin{aligned} \ell(W) = \ell_{X,Y}(W) = \|WX - Y\|^2 &= \langle Y, Y \rangle - 2\langle WX, Y \rangle + \langle WX, WX \rangle \\ &= \mathrm{const.} + 2\langle W(XX^\top), YX^\top(XX^\top)^{-1} \rangle + \langle WX, WX \rangle \\ &= \mathrm{const.} + 2\langle W, U \rangle_{XX^\top} + \langle W, W \rangle_{XX^\top} \\ &= \mathrm{const.} + \|W - U\|^2_{XX^\top}, \end{aligned}$$

(6.1)

where U is the unconstrained optimizer of this quadratic problem, over the set of all matrices:

$$U = \operatorname*{argmin}_{V \in \mathbb{R}^{d_L \times d_0}} \|VX - Y\|^2 = YX^\top(XX^\top)^{-1}.$$

As discussed in the previous sections, an LCN corresponds to a semi-algebraic set $\mathcal{M}$ of convolutional matrices $\overline{W}$. By (6.1), the optimization of the square error $\ell(\overline{W})$ over the function space $\mathcal{M}$ corresponds to a constrained linear regression problem, which is equivalent to solving

$$\min_{\overline{W} \in \mathcal{M}} \|U - \overline{W}\|^2_{XX^\top}.$$

If $U \in \mathcal{M}$, then the solution is U. If $U \notin \mathcal{M}$, we can optimize the objective over the boundary of the function space. We also may assume that the matrix U is convolutional of the same format as the matrices in $\mathcal{M}$, for if not, we can consider the orthogonal projection (with respect to $\|\cdot\|_{XX^\top}$) to the appropriate linear space of convolutional matrices.

Remark 6.1. The same expressions as above hold for population losses if we replace $\frac{1}{N} XX^\top = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^\top$ with $\mathbb{E}_{x \sim \mathcal{D}}[xx^\top]$ where $\mathcal{D}$ is the distribution over the input data.

If we assume that U is a convolutional matrix with filter u, then we can also write the square loss in terms of the filter $\overline{w}$ of $\overline{W}$:

$$
\ell(\overline{w}) = \|\overline{W} - U\|_{XX^\top}^2 = \operatorname{tr}[(\overline{W} - U)XX^\top(\overline{W} - U)^\top]
$$
(6.2)
$$
= \sum_{m=0}^{d_L-1} [(\overline{W} - U)_{m,:}XX^\top(\overline{W} - U)_{m,:}^\top],
$$

where we write $M_{m,:}$ for the mth row of a given matrix M. We denote by k and s the filter size and stride of the convolutional matrices in the function space $\mathcal{M}$. Since the mth row of $\overline{W} - U$ is obtained by shifting $\overline{w} - u$ by sm positions to the right, we see that

$$
(6.3) \qquad\qquad \ell(\overline{w}) = (\overline{w} - u)^\top \tau(XX^\top)(\overline{w} - u),
$$

where $\tau : \mathbb{R}^{d_0 \times d_0} \to \mathbb{R}^{k \times k}$ is an operator that sums all shifts of rows and columns of a matrix:

$$
(6.4) \qquad\qquad \tau(M)_{ij} = \sum_{m=0}^{d_L-1} M_{i+sm,j+sm}.
$$

Remark 6.2. If we consider circulant matrices instead of convolutional matrices, then all matrices appearing in (6.2) would be square matrices of size $d_0 \times d_0$ and the trace would sum from $m = 0$ to $d_0 - 1$. The sum in (6.4) would change accordingly and would be cyclic.

Lemma 6.3. *Let $XX^\top$ be a multiple of the identity matrix. For any stride s, the square loss ℓ on the filters is equivalent to the Euclidean distance between $\overline{w}$ and u, i.e., $\ell(\overline{w}) \propto \|\overline{w} - u\|^2$.*

Proof. If $XX^\top$ is a multiple of the identity, then $\tau(XX^\top) = d_L XX^\top$ for convolutional matrices (resp., $\tau(XX^\top) = d_0 XX^\top$ for circulant matrices). Now the claim follows from (6.3). ∎

Remark 6.4. The lemma above holds both for convolutional and circulant matrices. If we restrict ourselves to the circulant setting, then the statement holds more generally if $XX^\top$ is an arbitrary diagonal matrix as long as s and d_0 are coprime. This follows from Remark 6.2. Indeed, if s and d_0 are coprime, $\tau(XX^\top)_{ii}$ sums over all entries of the diagonal matrix $XX^\top$, and so $\tau(XX^\top)$ is a multiple of the identity matrix.

We can apply Lemma 6.3 also for population losses (see Remark 6.1) where the data distribution $\mathcal{D}$ is such that the covariance matrix $\mathbb{E}_{x \sim \mathcal{D}}[xx^\top]$ is a multiple of the identity (e.g., Gaussian data) and conclude that using the square loss on the LCN is equivalent to using a standard Euclidean distance at the level of filters. If we make use of circulant convolutions and restrict ourselves to a stride that is coprime to d_0, then this is also the case if the covariance matrix is simply diagonal.

Next we give a detailed example illustrating the critical points of the quadratic loss for an LCN and how these depend on the particular training data. In the next subsection, we will discuss the number of critical points in more generality.

Example 6.5. We return to Example 2.4 to consider minimizing the square loss over the set of functions $\mathcal{M} = \{W = (A, B, C) \colon B^2 \geq 4AC\}$ represented by the two-layer LCN with architecture specified by $\boldsymbol{d} = (3, 2, 1)$, $\boldsymbol{k} = (2, 2)$, $\boldsymbol{s} = (1, 1)$. Given data matrices X, Y, we are minimizing an objective $\ell(W) = \|WX - Y\|^2$ which is convex over the space of all matrices $W \in \mathbb{R}^{1 \times 3}$. Hence, if the unconstrained minimizer is outside $\mathcal{M}$ and $XX^\top$ is full rank, then the set of minimizers over $\mathcal{M}$ will be at the boundary of $\mathcal{M}$. The situation is illustrated in Figure 1. The boundary $\partial\mathcal{M} = \{W = (A, B, C) \colon B^2 = 4AC\}$ can be parameterized by $(A, C) \mapsto W = (A, \pm 2\sqrt{AC}, C)$ with (A, C) either both nonnegative or both nonpositive. In the following, we consider the case $A, C \geq 0$ (the case $A, C \leq 0$ is similar). The gradient of the objective function with respect to A, C is

$$\nabla_{A,C}\ell(A, C) = (-2YX^\top + 2WXX^\top)\nabla_{A,C}W^\top.$$

The first-order criticality condition $\nabla_{A,C}\ell(A, C) = 0$ means that

$$(6.5) \qquad\qquad -YX^\top + WXX^\top = \lambda WJ,$$

where we introduce the matrix $J = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1/2 & 0 \\ 1 & 0 & 0 \end{bmatrix}$, so that $WJ = [C, \mp\sqrt{AC}, A]$ is a basis of the left null space of $\nabla_{A,C}W^\top$. We can solve (6.5) by writing

$$(6.6) \qquad\qquad W = YX^\top(XX^\top - \lambda J)^{-1}$$

and plugging this into the defining equation of the boundary, $B^2 = 4AC$, which gives us an equation in λ. As discussed in [28, section 4.4], computing a square distance from a point to a quadric surface results in a degree 6 polynomial in λ. Since our boundary equation is homogeneous, we can clear the determinant in Cramer's rule for the matrix inverse and use simply the adjugate matrix, which results in a degree 4 polynomial in λ:

$$(6.7) \quad (YX^\top \operatorname{adj}(XX^\top - \lambda J)_{:,2})^2 - 4(YX^\top \operatorname{adj}(XX^\top - \lambda J)_{:,1})(YX^\top \operatorname{adj}(XX^\top - \lambda J)_{:,3}).$$

Solving this, we typically obtain four values of λ, which when inserted in (6.6) give us the critical points in the boundary of the function space.

Now we take a closer look at the special case with $XX^\top = I$. In this case, the unconstrained optimal solution is $U = (u_1, u_2, u_3) = YX^\top$ and (6.7) simplifies to

$$(6.8) \quad (u_2^2 - u_1u_3)\lambda^4 + (-u_1^2 - 4u_1u_3 - u_3^2)\lambda^3 + (-4u_1^2 - 2u_2^2 - 5u_1u_3 - 4u_3^2)\lambda^2$$
$$+ (-4u_1^2 - 4u_1u_3 - 4u_3^2)\lambda + (u_2^2 - 4u_1u_3).$$

We can classify the points (u_1, u_2, u_3) for which the problem has different numbers of solutions. When $u_2^2 - u_1u_3 \neq 0$, (6.8) is a degree 4 polynomial in λ which generally has four or two real
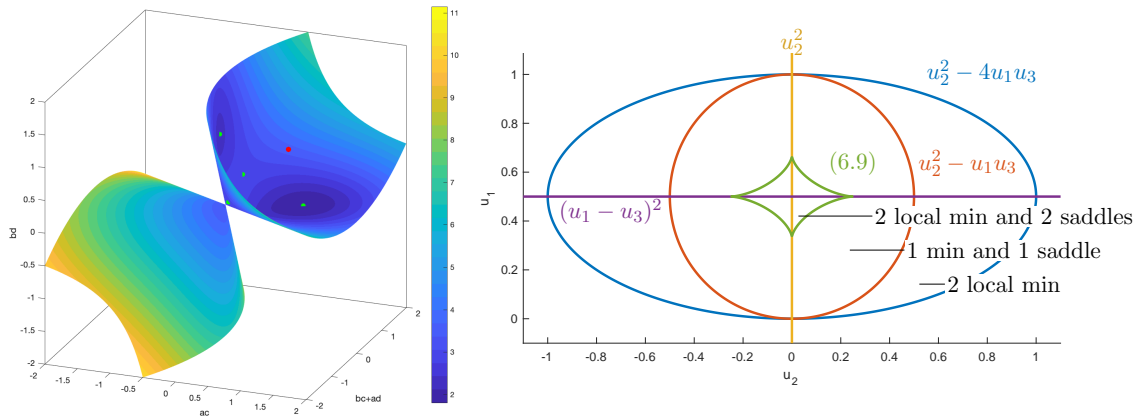
Figure 1. *Left: Boundary of the function space in $\mathbb{R}^{1\times 3}$ from* Example *2.4. The complement of the function space is inside the elliptic double cone with axis $(1,0,1)$. The boundary is colored by values of the square loss for random data X,Y with $U = YX^\top(XX^\top)^{-1}$ shown as a red dot. For this data, there are two local minima and two saddles shown as green dots. Right: Discrimination of datasets with $XX^\top = I$ in terms of $U = (u_1, u_2, u_3) = YX^\top$ by the types of critical points of the square loss as discussed in* Example *6.5. Shown is the slice $u_3 = -u_1 + 1$. For points with $u_2^2 - 4u_1u_3 < 0$ outside the function space, we have cases with two local minima and two saddles, one minimum and one saddle, and two local minima.*

solutions (the global minimum corresponds to a real root). The cases are separated by the discriminant of (6.8) with factors u_2^2, $(u_1 - u_3)^2$, $(u_1 + u_3)^2$, and

$$(6.9) \quad 32u_1^6 + 435u_1^4u_2^2 + 384u_1^2u_2^4 + 256u_2^6 - 240u_1^5u_3 - 960u_1^3u_2^2u_3 - 960u_1u_2^4u_3 + 696u_1^4u_3^2$$
$$+ 1098u_1^2u_2^2u_3^2 + 384u_2^4u_3^2 - 980u_1^3u_3^3 - 960u_1u_2^2u_3^3 + 696u_1^2u_3^4 + 435u_2^2u_3^4 - 240u_1u_3^5 + 32u_3^6.$$

We visualize the situation in Figure 1. The figure shows the slice $u_3 = -u_1 + 1$. Since the discriminant is homogeneous, each point in the slice is representative of a line through the origin in $\mathbb{R}^3$. The zeros of the factor $(u_1 + u_3)^2$ are not visible over this slice. We see that the discriminant separates 12 generic cases outside of the function space (i.e., inside the blue ellipse, where $u_2^2 - 4u_1u_3 < 0$). By evaluating individual instances in each of these cases, we find that the points inside the green curvy diamond (caustic) have two local minima and two saddles, the points between the green curvy diamond and the red circle have one minimum and one saddle, and the points between the red circle and the blue ellipse have two local minima (whereby we observed that only one of them satisfies $A, C \geq 0$).

6.2. Euclidean distance degrees. We now discuss the number of critical points of the optimization problem in function space using the square loss (6.1). This problem is given by

$$\min_{W \in \mathcal{M}} \|U - W\|_{XX^\top}^2,$$

where $\mathcal{M}$ is the function space of an LCN and U and X are fixed data matrices. We focus on one-dimensional convolutions with stride one. We first assume that $XX^\top$ is the identity matrix (i.e., we are interested in minimizing the standard Euclidean distance of the data U to the function space $\mathcal{M}$) and comment on other choices of X later.

By Theorem 5.2, each critical point in the parameter space of the LCN corresponds in function space to a critical point of the Euclidean distance from the data U to a multiple root locus Δ_λ. For almost every choice of U, the number of *complex* critical points of minimizing the Euclidean distance from U to Δ_λ is the same. This number is known as the *Euclidean distance degree* (*ED degree* for short) of Δ_λ [15]. Hence, if we assume that the data U is generic, Corollary 5.5 shows that the number of $\mu(\theta)$, where θ is a critical point in the parameter space of the given LCN, is upper bounded by the sum of the ED degrees of all multiple root loci $\Delta_\lambda \subset \mathbb{R}[\mathsf{x}, \mathsf{y}]_{k-1}$, where k is the filter size in the function space $\mathcal{M}$ and $\lambda = \lambda_{\rho|\gamma}$ comes from an rrmp $(\rho \mid \gamma)$ that is compatible with the LCN architecture.

Example 6.6. For generic data U, Example 5.6 shows the following: The number of $\mu(\theta)$, where θ is a critical point of an LCN with architecture $\boldsymbol{k} = (3, 2, 2)$ (resp., $\boldsymbol{k'} = (4, 2)$), is at most the sum of the ED degrees of $\Delta_{(1,1,1,1)}, \Delta_{(1,1,2)}, \Delta_{(2,2)}, \Delta_{(1,3)}$ (resp., $\Delta_{(1,1,1,1)}, \Delta_{(1,1,2)}$).

The state-of-the-art on ED degrees of multiple root loci is presented in [27]. For instance, the ED degree of the discriminant hypersurface of degree-$(k-1)$ polynomials is $3(k-1) - 2$ [15, Example 5.12]. More generally, the ED degree of the multiple root locus $\Delta_{(\alpha,1,\dots,1)}$ (with $\alpha > 1$ and $k - 1 - \alpha$ ones) is $(2\alpha - 1)(k - 1) - 2(\alpha - 1)^2$ [27, Theorem 5.1]. Many more ED degrees are computed in [27, Table 1], but a closed formula for general partitions is not known.

Example 6.7 ($k = 5$). The discriminant of quartic polynomials has ED degree 10. Its iterated singular loci are the following multiple root loci: $\Delta_{(2,2)}$ with ED degree 13, $\Delta_{(1,3)}$ with ED degree 12, and $\Delta_{(4)}$ with ED degree 10. Hence, the upper bounds on the number of $\mu(\theta)$ for critical points θ given in Example 6.6 are $1 + 10 + 13 + 12 = 36$ (resp., $1 + 10 = 11$).

Remark 6.8. The ED degrees discussed in this section are often referred to as *generic ED degrees* in the algebraic literature. The *special ED degree* refers to the following slightly different norm to be optimized: for two polynomials $P = \sum_{i=0}^{k-1} w_i \mathsf{x}^i$ and $Q = \sum_{i=0}^{k-1} u_i \mathsf{x}^i$,

$$(6.10) \qquad \langle P, Q \rangle_B := \sum_{i=0}^{k-1} \frac{i!(k-i-1)!}{(k-1)!} u_i w_i.$$

The letter "B" stands for *Bombieri norm*. For example, if P and Q are quadratic polynomials, then $\langle P, Q \rangle_B = u_2 w_2 + \frac{1}{2} u_1 w_1 + u_0 w_0$. The Euclidean inner product that arises for LCNs if $XX^\top$ is the identity is instead simply $\langle P, Q \rangle_2 = u_2 w_2 + u_1 w_1 + u_0 w_0$. This slight difference has an impact on the ED degrees. Indeed, the ED degree of the discriminant hypersurface $\{(w_0, w_1, w_2): w_1^2 = w_0 w_2\}$ is 2 for the Bombieri norm and 4 for the standard Euclidean norm (see Example 6.5 for a computation of the four critical points). More generally, if we consider the multiple root loci for polynomials of degree $k - 1$, then the special ED degree of the discriminant hypersurface is $k - 1$ [15, Corollary 8.7] (whereas the generic ED degree is $3(k-1) - 2$, as noted above). The special ED degree of the multiple root locus $\Delta_{(\alpha,1,\dots,1)}$ (with $\alpha > 1$ and $k - 1 - \alpha$ ones) is $k - 1$, independently of α [27, Theorem 5.1].

Example 6.9 ($k = 5$). The generic ED degrees of the multiple root loci of quartic polynomials are listed in Example 6.7. The special ED degrees are smaller: for $\Delta_{(1,1,2)}$ it is 4, for $\Delta_{(2,2)}$ it is 7, for $\Delta_{(1,3)}$ it is 4, and for $\Delta_{(4)}$ it is 4.

Remark 6.10. The above discussion shows that the Bombieri norm can be thought of as a "special" norm that leads to lower ED degrees and thus makes the optimization landscape simpler. Perhaps counterintuitively, the standard Euclidean distance is instead *generic*, in the sense that almost all pairings $\langle \cdot, \cdot \rangle$ will lead to the same ED degrees for multiple-root loci. In other words, for *almost all* data matrices the ED degrees of the multiple root loci would be the same as assuming $XX^\top$ is the identity.

A natural question in this setting is if there exists a "special" data distribution that induces the Bombieri norm (6.10) in polynomial space. For circulant convolutions, this is not possible. Indeed, for any matrix M, the matrix $\tau(M)$ when $d_0 = d_L$ will always have constant diagonal entries by the same reasoning as in Remark 6.4. In the convolutional (Toeplitz) case, however, such special data distributions can exist, as the following simple example demonstrates.

Example 6.11. We consider a single convolutional layer with $k = 3$, $d_0 = 4$, and $d_1 = 2$. If the data covariance matrix is such that

$$XX^\top = \begin{bmatrix} 3/4 & & & \\ & 1/4 & & \\ & & 1/4 & \\ & & & 3/4 \end{bmatrix},$$

then the induced metric in the space of quadratic polynomials is indeed the Bombieri norm $\langle P, Q \rangle_B = u_2 w_2 + \frac{1}{2} u_1 w_1 + u_0 w_0$. This example suggests that nonstandard normalizations of the data might lead to more favorable optimization landscapes for convolutional networks.

Example 6.12. We return to Example 6.5, an LCN with $\mathcal{M} = \{W = (A, B, C) \colon B^2 \geq 4AC\}$, and investigate the structure of critical points of the square loss over the boundary $\partial\mathcal{M} = \{W = (A, B, C) \colon B^2 = 4AC\}$ when $XX^\top$ is compatible with the Bombieri norm,

$$(6.11) \qquad XX^\top = \begin{bmatrix} 1 & & \\ & 1/2 & \\ & & 1 \end{bmatrix}.$$

A geometric interpretation of this setting is that the balls $\{W \colon \|W - U\|_{XX^\top} \leq c\}$ are ellipsoids with the same $1 \times 2 \times 1$ aspect ratio as the boundary of the function space. As before, we write $U = (u_1, u_2, u_3) = YX^\top(XX^\top)^{-1}$ for the unconstrained least squares solution. With $XX^\top$ taking the form (6.11), equation (6.7) for the case $A, C \geq 0$ becomes

$$(6.12) \qquad (-1)(\lambda + 1)(\lambda + 1)((u_1 u_3 - 4u_2^2)\lambda^2 + (u_1^2 + 8u_2^2 + u_3^2)\lambda + (u_1 u_3 - 4u_2^2)).$$

This quartic polynomial in λ has a repeated root $\lambda = -1$ for any values of U, and hence its discriminant vanishes identically. The distinction in terms of U is determined by the last factor alone. We discuss the zeros of the individual factors in turn.

- Consider the factor $(\lambda + 1)$ of (6.12). The root $\lambda = -1$ plugged into (6.5) gives equation $W(XX^\top + J) = YX^\top$ for W. Since the matrix $XX^\top + J = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ is not of full rank, for generic $YX^\top$ there is no solution and this case can be ignored. Solutions only exist in the nongeneric case that $YX^\top = c \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}$, namely $\{W = (A, B, C) \colon A + C = c\}$, which is a hyperplane. The intersection of this hyperplane with the boundary of the function space gives us the set of critical points $\{(A, B, C) \colon A + C = c, B^2 = 4AC\}$. For the assumed case $A, C \geq 0$, this is empty over the reals for $c < 0$, a single point for $c = 0$, and an ellipse for $c > 0$. All critical points have the same Bombieri distance c from $U = \begin{bmatrix} c & 0 & c \end{bmatrix}$.

- Consider now the last factor of (6.12), which is quadratic in λ. Its discriminant has factors $(u_1 + u_3)^2$ and $u_1^2 - 2u_1u_3 + 16u_2^2 + u_3^2$. The former has zero set $\{u_1 = -u_3\}$, which also appeared in the case $XX^\top = I$ and is inside the function space. The latter has real zero set $\{u_2 = 0, u_1 = u_3\}$, which is the center line of our boundary cone $\partial\mathcal{M}$ and can be interpreted as a degeneration of the caustic that appeared in the case $XX^\top = I$. It also corresponds to the nongeneric case discussed above.

Hence, when $XX^\top$ takes the form (6.11) compatible with the Bombieri norm, there is a single generic case for the datasets. We find that generic datasets $U = (u_1, u_2, u_3)$ outside the function space have two critical points over the boundary of the function space: one global minimum and one saddle. This is in contrast with the more complicated case distinction in Example 6.5 and Figure 1 when $XX^\top = I$, where some generic datasets led to four critical points. On the other hand, for the Bombieri norm there are special datasets, with $\{u_2 = 0, u_1 = u_3 > 0\}$, with infinitely many critical points, namely entire ellipses of global minima.

7. Examples and numerical evaluation. We present computational experiments on LCNs with one-dimensional convolutions with stride one. In Proposition 5.1, we saw that critical parameters are global minimizers of the objective or are contained in the discriminant hypersurface. In fact, in Theorem 5.2 we observed that any critical parameter corresponds to a critical point over the locus with the same root structure. We discuss the relevant root loci and evaluate their appearance in gradient optimization using different LCN architectures. As observed in the previous sections, the order of the filters does not affect the function space or the gradient dynamics. For this reason, we consider architectures with filter widths $k_1 \geq \cdots \geq k_L$.

We perform our experiments with the five nonfilling architectures from Example 4.6. The left-hand side of Figure 1 visualizes the function space $\mathcal{M}_{(2,2)}$. The function spaces representing cubic and quartic polynomials are depicted in Figure 2: On the left, $\mathcal{M}_{(2,2,2)}$ is the "inside" of the "cuspidal wedge." On the right, the complement of $\mathcal{M}_{(3,2,2)} = \mathcal{M}_{(4,2)}$ are the points above the blue "bowl." These points correspond to one of the two convex cones described in Proposition 4.7. The other cone is not visible since the figure fixes the highest degree coefficient to 1. Moreover, $\mathcal{M}_{(2,2,2,2)}$ is the set inside the triangular region.

7.1. Critical points reached by gradient descent. For each of the architectures listed above, we generated $10,000$ random datasets X, Y, each consisting of 10 i.i.d. input-output pairs from a standard normal distribution with inputs in $\mathbb{R}^{d_0}$ and outputs in $\mathbb{R}$. Since $d_L = 1$, we have from Proposition 2.2 that $d_0 = k$. For each dataset and architecture, we initialized the parameters at random from a standard normal distribution and ran gradient descent optimization of the square loss $\mathcal{L}$ with fixed step size 0.01 and stopping criterion $\|\nabla\mathcal{L}\|^2 \leq 10^{-14}$. Instances which did not reach the stopping criterion within a maximum of $15,000$ iterations were discarded. Since the final coefficients are only approximately critical points, they rarely correspond to polynomials with exactly repeated roots. To assess the rrmp's of solutions, we compute the roots of all factor filters and consider a root r as real if $|r - \text{real}(r)| \leq \texttt{tol} \cdot |r|$, and we consider two roots r, r' the same if $|r - r'| \leq \texttt{tol} \cdot \max\{|r|, |r'|\}$ with $\texttt{tol} = 10^{-4}$. This classification scheme is relatively robust but is not exact. The results are shown in Table 1.
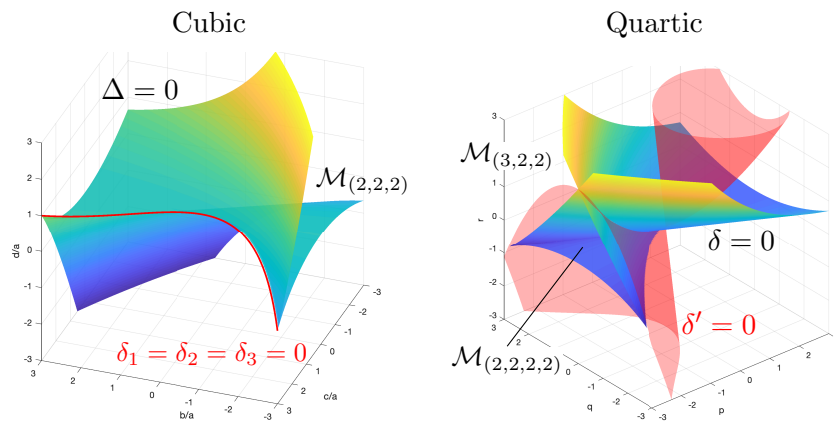
Figure 2. *Left: The surface $\Delta = 0$ of coefficients (a, b, c, d), $a \neq 0$, for which a cubic polynomial $a\mathsf{x}^3 + b\mathsf{x}^2\mathsf{y} + c\mathsf{x}\mathsf{y}^2 + d\mathsf{y}^3$ has a double root. The coefficients with triple roots are shown in red. The coefficients with $\Delta > 0$ (the interior of $\mathcal{M}_{(2,2,2)}$) are on the side with smaller values of c. Right: The surfaces $\delta = 0$ and $\delta' = 0$ (as in Example 3.3) of coefficients (p, q, r) that separate the real root multiplicities of a depressed quartic polynomial $\mathsf{x}^4 + p\mathsf{x}^2\mathsf{y}^2 + q\mathsf{x}\mathsf{y}^3 + r\mathsf{y}^4$. The positive values of δ and δ' are on the sides that do not contain the labels. The complement of $\mathcal{M}_{(3,2,2)}$ is above the blue bowl. The interior of $\mathcal{M}_{(2,2,2,2)}$ is within the triangular blue region.*

- For the architecture $\boldsymbol{k} = (2, 2)$, about 66% of the datasets correspond to targets with rrmp $11|0$ inside the function space, and gradient descent finds a solution with nearly zero training error. About 34% of the datasets correspond to targets with rrmp $0|1$ outside the function space, and gradient descent finds solutions with rrmp $2|0$ at its boundary.

- For $\boldsymbol{k} = (2, 2, 2)$, about 25% of the datasets have rrmp $111|0$ inside the function space, and for most of these, gradient descent finds a solution of the same rrmp with nearly zero loss. About 75% of the datasets have rrmp $1|1$ outside the function space. Of these, about 70% led to a solution with rrmp $12|0$, i.e., a smooth point of the discriminant hypersurface, and about 30% led to a solution with rrmp $3|0$, i.e., a point on the red curve in Figure 2.

- For the architectures $\boldsymbol{k} = (4, 2)$ or $(3, 2, 2)$, which have the same function space, the different parameterizations lead to different optimization behaviors. First, the initialization procedure leads to different distributions in function space, with $\boldsymbol{k} = (4, 2)$ being more frequently initialized with rrmp $11|1$ than $\boldsymbol{k} = (3, 2, 2)$. For targets with rrmp $0|11$, outside the function space, the solutions often correspond to zeros of the discriminant and are critical points in parameter space, but they do not lie on the Euclidean boundary of $\mathcal{M}_{\boldsymbol{k}}$. Hence, they are spurious critical points. The effect is more pronounced in the deeper architecture, which has a more intricate parameterization. The optimization landscapes are illustrated in Figure 3. For targets with rrmp $11|1$, inside the function space, a good fraction of the optimization instances with $\boldsymbol{k} = (4, 2)$ converged to suboptimal solutions with rrmp $112|0$ or $2|1$, and for $\boldsymbol{k} = (3, 2, 2)$ an even larger fraction converged to suboptimal solutions with rrmp $112|0$, $22|0$, $13|0$,

or 2|1. This is in line with our theoretical discussion, by which $\boldsymbol{k} = (3, 2, 2)$, but not $\boldsymbol{k} = (4, 2)$, can have critical points in parameter space with rrmp's 22|0 and 13|0 for generic data (see Corollary 5.5 and Example 5.6).

- For $\boldsymbol{k} = (2, 2, 2, 2)$, in contrast to the previous architectures, a fraction of the targets outside the function space (i.e., with rrmp 11|1 or 0|11) led to a solution with rrmp 4|0, i.e., the codimension 3 point at the origin in Figure 2, which was predicted by Corollary 5.5.

7.2. Euclidean vs. Bombieri norm. In this experiment, we generated 500 random targets U for each architecture. For each of them, we ran gradient descent with 50 random initializations for the loss with Euclidean and Bombieri norms. For each target, we recorded the list of converged solutions using the settings of the previous experiment. To determine the number of distinct solutions, we considered two filters w_i and $w_{i'}$ as equal if on each entry $|w_{ik} - w_{i'k}| \leq \mathtt{tol} \cdot \max\{\max\{|w_{jk}| : j\}, 1\}$ with $\mathtt{tol} = 10^{-4}$. The results are shown in Table 2. In line with our theoretical discussion in subsection 6.2, we observed fewer distinct solutions for the Bombieri norm than for the generic norm.

7.3. Symbolic computations. We illustrate several of our results with a final example. We consider linearly generated data $\mathcal{D}_N = \{(x_i, y_i) \in \mathbb{R}^5 \times \mathbb{R} : i = 1, \ldots, N, y_i = u^\top x_i\}$, where $u = [2, 0, 5, 0, 2]^\top \in \mathbb{R}^5$ is a "target filter" and x_i is sampled from a standard Gaussian distribution. We study the square loss regression problem for different one-dimensional LCN architectures with compatible dimensions, as described in section 6. In light of Lemma 6.3, if the data is normalized so that the empirical covariance $XX^\top$ is the identity, or if $N \to \infty$, then this corresponds to an objective function $\mathcal{L}(\theta) = \ell(\mu(\theta)) = \|\mu(\theta) - u\|^2$, where $\overline{w} = \mu(\theta)$ is the end-to-end filter and $\|\cdot\|$ is the standard Euclidean norm.

The possible LCN architectures with stride one, $d_0 = 5, d_L = 1$, and nontrivial filter sizes $1 < k_i < 5$ are $\boldsymbol{k} = (4, 2), (3, 3), (3, 2, 2)$, and $(2, 2, 2, 2)$ (not considering permutations). By Theorem 4.1, only the architecture $\boldsymbol{k} = (3, 3)$ is filling. The target filter corresponds to the polynomial $\pi(u) = 2\mathsf{x}^4 + 5\mathsf{x}^2\mathsf{y}^2 + 2\mathsf{y}^4 = (\mathsf{x}^2 + 2\mathsf{y}^2)(2\mathsf{x}^2 + \mathsf{y}^2)$, which does not have any real roots and is thus not contained in $\mathcal{M}_{(4,2)}$, $\mathcal{M}_{(3,2,2)}$, or $\mathcal{M}_{(2,2,2,2)}$.

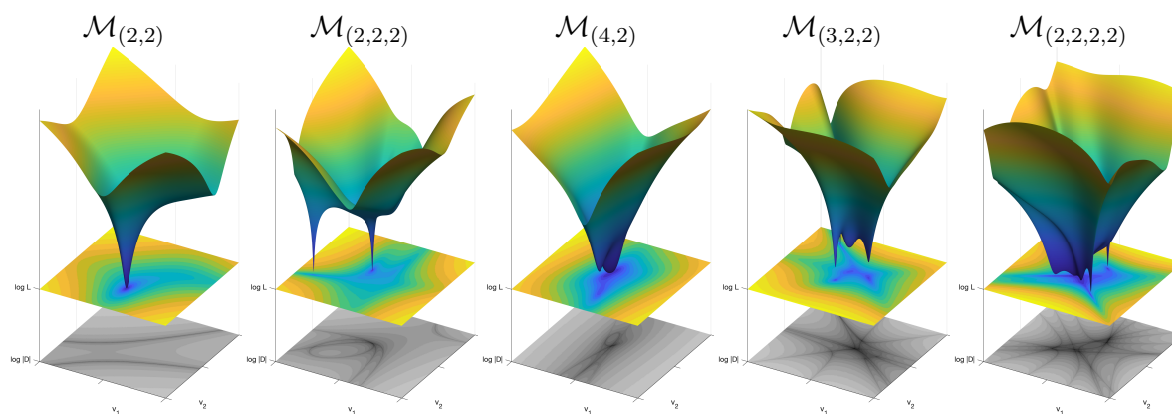Following our results in subsection 5.1, for all architectures listed above, any critical point $\theta \in \mathrm{Crit}(\mathcal{L})$ with $\mu(\theta) \neq u$ corresponds to polynomials with repeated roots. More precisely, if θ is a critical point for $\mathcal{L}$, then $\mu(\theta)$ is a critical point for $\ell|_{\Delta_\lambda}$ where $\ell(\overline{w}) = \|\overline{w} - u\|^2$ and Δ_λ is the multiple root locus where λ describes the (complex) root pattern of $\mu(\theta)$. Based on this, we can compute the critical points for $\ell|_{\Delta_\lambda}$ for any λ by solving algebraic systems. Recall that the number of complex critical points was predicted by the ED degrees in Example 6.7.

- $\lambda = (2, 1, 1)$: Out of the 10 complex critical points on Δ_λ, four are real and in fact rational, namely $[\frac{1}{5}, \pm\frac{9}{5}, \frac{16}{5}, \pm\frac{9}{5}, \frac{1}{5}]^\top \equiv \frac{1}{5}(\mathsf{x} \pm \mathsf{y})^2(\mathsf{x}^2 \pm 7\mathsf{x}\mathsf{y} + \mathsf{y}^2)$, $[0, 0, 5, 0, 2]^\top \equiv \mathsf{y}^2(5\mathsf{x}^2 + 2\mathsf{y}^2)$, and $[2, 0, 5, 0, 0]^\top \equiv \mathsf{x}^2(5\mathsf{y}^2 + 2\mathsf{x}^2)$.

- $\lambda = (3, 1)$: Out of the 12 complex critical points on Δ_λ, four are real and none are rational.

- $\lambda = (2, 2)$: Out of the 13 complex critical points on Δ_λ, five are real and three are rational: $[-1, 0, 2, 0, -1]^\top \equiv -(\mathsf{x} - \mathsf{y})^2(\mathsf{x} + \mathsf{y})^2$, $[0, 0, 5, 0, 0]^\top \equiv 5\mathsf{x}^2\mathsf{y}^2$, $[\frac{7}{3}, 0, \frac{14}{3}, 0, \frac{7}{3}]^\top \equiv \frac{7}{3}(\mathsf{x}^2 + \mathsf{y}^2)^2$.

- $\lambda = (4)$: Out of the 10 complex critical points on Δ_λ, four are real and also rational: $[0, 0, 0, 0, 2]^\top \equiv 2\mathsf{y}^4$, $[2, 0, 0, 0, 0]^\top \equiv 2\mathsf{x}^4$, $[\frac{17}{35}, \pm\frac{68}{35}, \frac{102}{35}, \pm\frac{68}{35}, \frac{17}{35}]^\top \equiv \frac{17}{35}(\mathsf{x} \pm \mathsf{y})^4$.

Table 1

Rrmp's of solutions returned by gradient descent optimization of the square loss in parameter space for random datasets and initializations. The table omits a small fraction of targets and initializations numerically classified in rrmp's of positive codimension.

$\mathcal{M}_{(2,2)}$

target	%	initialization 11|0 100%		
		solution	%	mean loss
11|0	65.8	11|0	100	1.15e-15
0|1	34.2	2|0	100	0.185

$\mathcal{M}_{(2,2,2)}$

target	%	initialization 111|0 100%		
		solution	%	mean loss
111|0	24.9	111|0	99.7	2.32e-15
		12|0	0.051	0.945
		3|0	0.205	1.19
1|1	75	12|0	69.8	0.224
		3|0	30.2	0.709

$\mathcal{M}_{(4,2)}$

target	%	initialization 1111|0 24.6%			initialization 11|1 75.3%		
		solution	%	mean loss	solution	%	mean loss
1111|0	5.28	1111|0	100	3.04e-15	1111|0	100	3.1e-15
11|1	72.6	112|0	15.5	0.228	112|0	12.4	0.193
		11|1	83.2	1.94e-15	11|1	86.7	1.84e-15
		2|1	1.36	0.54	2|1	0.886	0.397
0|11	22.1	112|0	7.85	0.347	112|0	4.81	0.35
		2|1	92.2	0.231	2|1	95.2	0.229

$\mathcal{M}_{(3,2,2)}$

target	%	initialization 1111|0 65.1%			initialization 11|1 34.9%		
		solution	%	mean loss	solution	%	mean loss
1111|0	4.82	1111|0	99.6	4.68e-15	1111|0	100	4e-15
		13|0	0.429	0.71			
11|1	72.9	112|0	27.1	0.221	112|0	21.8	0.207
		22|0	1.28	0.992	22|0	0.663	0.879
		13|0	25.8	0.798	13|0	15.6	0.71
		11|1	45.5	1.78e-15	11|1	61.7	1.7e-15
		2|1	0.381	0.446	2|1	0.306	0.418
0|11	22.3	112|0	11.2	0.374	112|0	9.09	0.365
		22|0	25.5	0.855	22|0	17.5	0.882
		13|0	7.1	0.895	13|0	4.21	0.937
		2|1	56.2	0.224	2|1	69.2	0.208

$\mathcal{M}_{(2,2,2,2)}$

target	%	initialization 1111|0 99.8%		
		solution	%	mean loss
1111|0	4.79	1111|0	99.7	5.41e-15
		13|0	0.274	0.297
11|1	72.7	112|0	33.3	0.259
		22|0	2.24	0.908
		13|0	52.1	0.774
		4|0	12.4	1.53
0|11	22.5	112|0	13.2	0.411
		22|0	40.7	0.779
		13|0	16.1	1.02
		4|0	30	1.48

$\mathcal{M}_{(2,2)}$ $\mathcal{M}_{(2,2,2)}$ $\mathcal{M}_{(4,2)}$ $\mathcal{M}_{(3,2,2)}$ $\mathcal{M}_{(2,2,2,2)}$

Figure 3. *The square loss* $(w_1, \ldots, w_L) \mapsto \|U - \mu(w_1, \ldots, w_L)\|^2_{XX^\top}$, $U = YX^\top(XX^\top)^{-1}$, *for random data* X, Y, *plotted in log-scale over random two-dimensional affine subsets of parameter values. The bottom shows the discriminant at the corresponding filters,* $(w_1, \ldots, w_L) \mapsto |\Delta(\pi(\mu(w_1, \ldots, w_L)))|$. *As can be seen, deeper networks tend to have a more intricate loss surface and critical parameters often correspond to functions where the discriminant vanishes.*

Table 2

Percentages of targets for which gradient descent with different initializations converged to a given number of distinct solutions. For the Bombieri norm B, we usually observe a smaller number of distinct solutions compared with the generic Euclidean norm I.

		Number of distinct solutions			
$\boldsymbol{k}$	Metric	1	2	3	4
$(2, 2)$	I	95.8	4.2		
	B	100			
$(2, 2, 2)$	I	84.7	14.5	0.84	
	B	95.6	1.1	3.29	
$(4, 2)$	I	66.3	33.7		
	B	73.8	26.2		
$(3, 2, 2)$	I	22.5	58.8	17.7	1.01
	B	32	48.3	19.5	0.209
$(2, 2, 2, 2)$	I	70.7	26.5	2.61	0.217
	B	84.8	8.21	6.97	

For any choice of quartic architecture, all critical points are mapped to either u or one of the 17 real critical points listed above. We have verified numerically that gradient descent indeed always converges to one of these critical points. In fact, it always seems to converge to one of the rational critical points. For example, for $\boldsymbol{k} = (4, 2)$, gradient descent converges to either $[2, 0, 5, 0, 0]^\top$ or $[0, 0, 5, 0, 2]^\top$ which arise with $\lambda = (2, 1, 1)$. Note that in this case we do not expect critical points with $\lambda = (3, 1), (2, 2), (4)$ due to Example 5.6 and Corollary 5.5.

Finally, we remark that the previous list only describes the critical points in function space but not the actual parameters. However, as discussed in subsection 5.2, it is possible to recover these parameters up to a finite ambiguity given the initialization of gradient descent. For example, assume that $\boldsymbol{k} = (4, 2)$ and we initialize gradient descent at $\theta(0) = (w_1(0), w_2(0)) = ([1, 6, 11, 6], [4, 1])$, i.e., a factorization of the polynomial $(x + y)(x + 2y)(x + 3y) \cdot (4x + y)$.

Then, following the approach outlined in Corollary 5.14, we can derive that if the gradient descent path $\theta(t)$ converges to $\mu(\theta) = [2, 0, 5, 0, 0]^\top$ in function space, then in parameter space $\theta(t)$ converges to one of $(w_1, w_2) = (1/\kappa[2, 0, 5, 0]^\top, \kappa[1, 0]^\top)$ with $\kappa = \pm\sqrt{\frac{1}{2}\left(\sqrt{31445} - 177\right)} \approx \pm 0.4045867$. Numerically, we verify that gradient descent converges to these filters with $\kappa > 0$.

8. Conclusion. We presented a semi-algebraic description of the families of functions that can be represented by LCNs. Using the fact that compositions of convolutions correspond to polynomial multiplication, we showed that the function space can be described as sets of polynomials that admit certain factorizations. We then investigated the optimization of an objective function over such sets and described the critical points in function space and in parameter space. Unlike fully connected linear networks, convolutional linear networks can have nonglobal local optimizers in function space as well as spurious local optimizers in parameter space. In more detail, our analysis showed that all critical points that are not global minima correspond to polynomials with repeated roots, according to multiplicity patterns that depend on the network's architecture. Possible extensions of this work might include a more detailed study of the function space for linear convolutions in higher dimensions or with larger strides (which we discussed briefly in sections 4.3 and 4.2) and, naturally, carrying out a similar analysis for *nonlinear* convolutional networks. In particular, one might first consider the special case of polynomial activation functions since in this setting the function space is still a semi-algebraic set.

Appendix A. Gradients. For convenience, we give a description of the gradient for LCNs.

Proposition A.1. *Consider a loss $\ell\colon \mathbb{R}^{d_L \times d_0} \to \mathbb{R}$ on the space of matrices. For an LCN parameterization $\mu\colon (w_1, \ldots, w_L) \mapsto \overline{W}$ as in (2.5), denote the loss on the space of filters by $\mathcal{L} = \ell \circ \mu$. The gradient is*

$$(\nabla_{w_l}\mathcal{L}(w_1, \ldots, w_L))_\alpha = \sum_{\beta,\gamma}(\nabla_{w_l}W_l)_{\alpha,\beta,\gamma}(W_{l+1}^\top \cdots W_L^\top \nabla_{\overline{W}}\ell(\overline{W})W_1^\top \cdots W_{l-1}^\top)_{\beta,\gamma},$$

where $(\nabla_{w_l}W_l)_{\alpha,\beta,\gamma} = \partial_{w_{l,\alpha}}(W_l)_{\beta,\gamma}$. For the square loss $\ell(\overline{W}) = \|\overline{W}X - Y\|^2$ with data X, Y, $\nabla_{\overline{W}}\ell(\overline{W}) = \overline{W}XX^\top - YX^\top$.

The previous result describes the gradient of the loss $\mathcal{L}$ in terms of convolutional matrices W_l; however, the expression does not depend on the actual matrices but only on the filters involved. We can make this fact explicit in the case of stride one as follows.

Proposition A.2. *Let $\mathcal{L} = \ell \circ \mu$ be a loss function for an LCN with stride one, and let $\theta = (w_1, \ldots, w_L)$ denote the filters. For any vector $w \in \mathbb{R}^k$, let $\pi_{-1}(w) := w_0\mathsf{x}^{-k+1} + w_1\mathsf{x}^{-k+2} + \cdots + w_{k-2}\mathsf{x}^{-1} + w_{k-1}$ be the Laurent polynomial (polynomial with negative exponents) whose coefficients are the elements of w. For any $i = 1, \ldots, L$, the gradient $\nabla_{w_i}\mathcal{L}(\theta) \in \mathbb{R}^{k_i}$ is given by the coefficients of degree $(0, \ldots, k_i - 1)$ in*

$$\tilde{\pi}(\nabla\ell) \cdot \pi_{-1}(w_1) \cdots \pi_{-1}(w_{i-1}) \cdot \pi_{-1}(w_{i+1}) \cdots \pi_{-1}(w_L) \in \mathbb{R}[\mathsf{x}, \mathsf{x}^{-1}],$$

where $\tilde{\pi}(\nabla\ell)$ is the univariate polynomial representation of the gradient $\nabla\ell(\mu(\theta))$ obtained from (2.8) by dehomogenizing $\mathsf{y} = 1$.

Proof. The Laurent polynomial $\pi_{-1}(w)$ is a representation of the *adjoint filter* $w^* := (w_{k-1}, \ldots, w_0)$. Convolution with the adjoint filter w^* is the adjoint operator of the convolution with w for the standard Euclidean inner product. In terms of polynomials, this means that

$$\langle \tilde{\pi}(v), \tilde{\pi}(w_2) \cdot \tilde{\pi}(w_1) \rangle = \langle \tilde{\pi}(v) \cdot \pi_{-1}(w_2), \tilde{\pi}(w_1) \rangle$$

for any vectors v, w_1, w_2, where $\langle \cdot, \cdot \rangle$ is the coefficientwise inner product and $\cdot$ is the product operation for polynomials (missing coefficients are treated as zero. so we need not constrain the sizes of v, w_1, w_2). The claim now follows from the fact that $\nabla_{w_i} \mathcal{L}$ is such that

$$\begin{aligned}
\langle \tilde{\pi}(\nabla_{w_i} \mathcal{L}), \tilde{\pi}(\dot{w}_i) \rangle &= \langle \tilde{\pi}(\nabla \ell), \tilde{\pi}(w_1) \cdots \tilde{\pi}(\dot{w}_i) \cdots \tilde{\pi}(w_L) \rangle \\
&= \langle \tilde{\pi}(\nabla \ell) \cdot \pi_{-1}(w_1) \cdots \pi_{-1}(w_{i-1}) \cdot \pi_{-1}(w_{i+1}) \cdots \pi_{-1}(w_L), \tilde{\pi}(\dot{w}_i) \rangle. \quad \blacksquare
\end{aligned}$$

REFERENCES

[1] Z. ALLEN-ZHU, Y. LI, AND Z. SONG, *A convergence theory for deep learning via over-parameterization*, in Proceedings of the 36th International Conference on Machine Learning, Proc. Mach. Learn. Res. 97, K. Chaudhuri and R. Salakhutdinov, eds., 2019, pp. 242–252, http://proceedings.mlr.press/v97/allen-zhu19a.html.

[2] E. ALLMAN, H. BANOS, R. EVANS, S. HOSTEN, K. KUBJAS, D. LEMKE, J. RHODES, AND P. ZWIERNIK, *Maximum likelihood estimation of the latent class model through model boundary decomposition*, J. Algebr. Stat., 10 (2019), pp. 51–84, https://doi.org/10.18409/jas.v10i1.75.

[3] D. S. ARNON, *Geometric reasoning with logic and algebra*, Artif. Intell., 37 (1988), pp. 37–60, https://doi.org/10.1016/0004-3702(88)90049-5.

[4] S. ARORA, N. COHEN, N. GOLOWICH, AND W. HU, *A convergence analysis of gradient descent for deep linear neural networks*, in International Conference on Learning Representations, 2019, https://openreview.net/forum?id=SkMQg3C5K7.

[5] S. ARORA, N. COHEN, AND E. HAZAN, *On the optimization of deep networks: Implicit acceleration by overparameterization*, in Proceedings of the 35th International Conference on Machine Learning, Proc. Mach. Learn. Res. 80, Stockholmsmässan, Stockholm, Sweden, pp. 244–253, http://proceedings.mlr.press/v80/arora18a.html.

[6] B. BAH, H. RAUHUT, U. TERSTIEGE, AND M. WESTDICKENBERG, *Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers*, Inf. Inference, 11 (2022), pp. 307–353.

[7] P. BALDI, *Linear learning: Landscapes and algorithms*, in Advances in Neural Information Processing Systems, Vol. 1, D. S. Touretzky, ed., Morgan Kaufmann, San Francisco, CA, 1989, pp. 65–72, http://papers.nips.cc/paper/123-linear-learning-landscapes-and-algorithms.pdf.

[8] P. BALDI AND K. HORNIK, *Neural networks and principal component analysis: Learning from examples without local minima*, Neural Netw., 2 (1989), pp. 53–58, https://doi.org/10.1016/0893-6080(89)90014-2.

[9] P. BALDI AND K. HORNIK, *Learning in linear neural networks: A survey*, IEEE Trans. Neural Netw., 6 (1995), pp. 837–858, https://doi.org/10.1109/72.392248.

[10] S. BASU, R. POLLACK, AND M.-F. ROY, *Algorithms in Real Algebraic Geometry (Algorithms and Computation in Mathematics)*, Springer-Verlag, Berlin, Heidelberg, 2006.

[11] J. F. BLINN, *How to solve a cubic equation, part* 1: *The shape of the discriminant*, IEEE Comput. Graph. Appl., 26 (2006), pp. 84–93, https://doi.org/10.1109/MCG.2006.60.

[12] T. Ö. ÇELIK, A. JAMNESHAN, G. MONTÚFAR, B. STURMFELS, AND L. VENTURELLO, *Optimal transport to a variety*, in Mathematical Aspects of Computer and Information Sciences, D. Slamanig, E. Tsigaridas, and Z. Zafeirakopoulos, eds., Springer, Cham, 2020, pp. 364–381.

[13] T. Ö. ÇELIK, A. JAMNESHAN, G. MONTÚFAR, B. STURMFELS, AND L. VENTURELLO, *Wasserstein distance to independence models*, J. Symb. Comput., 104 (2021), pp. 855–873, https://doi.org/10.1016/j.jsc.2020.10.005.

[14] Y. CHI, Y. M. LU, AND Y. CHEN, *Nonconvex optimization meets low-rank matrix factorization: An overview*, IEEE Trans. Signal Process., 67 (2019), pp. 5239–5269, https://doi.org/10.1109/TSP.2019.2937282.

[15] J. DRAISMA, E. HOROBEŢ, G. OTTAVIANI, B. STURMFELS, AND R. R. THOMAS, *The Euclidean distance degree of an algebraic variety*, Found. Comput. Math., 16 (2016), pp. 99–149.

[16] S. S. DU, X. ZHAI, B. POCZOS, AND A. SINGH, *Gradient descent provably optimizes over-parameterized neural networks*, in International Conference on Learning Representations, 2019, https://openreview.net/forum?id=S1eK3i09YQ.

[17] Y. DUKLER, Q. GU, AND G. MONTÚFAR, *Optimization theory for ReLU neural networks trained with normalization layers*, in Proceedings of the 37th International Conference on Machine Learning, Proc. Mach. Learn. Res. 119, H. Daumé III and A. Singh, eds., 2020, pp. 2751–2760, http://proceedings.mlr.press/v119/dukler20a.html.

[18] E. GONZALEZ AND D. WEINBERG, *Root Configurations of Real Univariate Cubics and Quartics*, preprint, https://arxiv.org/abs/1511.07489, 2015.

[19] D. GRAYSON AND M. STILLMAN, *Macaulay 2, a System for Computation in Algebraic Geometry and Commutative Algebra*, http://faculty.math.illinois.edu/Macaulay2/.

[20] S. GUNASEKAR, J. D. LEE, D. SOUDRY, AND N. SREBRO, *Implicit bias of gradient descent on linear convolutional networks*, in Advances in Neural Information Processing Systems, Vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., Curran Associates, Inc., Red Hook, NY, 2018, pp. 9461–9471, http://papers.nips.cc/paper/8156-implicit-bias-of-gradient-descent-on-linear-convolutional-networks.pdf.

[21] M. HARDT AND T. MA, *Identity matters in deep learning*, in Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 2017, https://openreview.net/forum?id=ryxB0Rtxx.

[22] R. HARTSHORNE, *Algebraic Geometry*, Grad. Texts in Math. 52, Springer, New York, 2013.

[23] A. JACOT, F. GABRIEL, AND C. HONGLER, *Neural tangent kernel: Convergence and generalization in neural networks*, in Advances in Neural Information Processing Systems, Vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., Curran Associates, Inc., Red Hook, NY, 2018, https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

[24] K. KAWAGUCHI, *Deep learning without poor local minima*, in Advances in Neural Information Processing Systems, Vol. 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., Curran Associates, Inc., Red Hook, NY, 2016, pp. 586–594, http://papers.nips.cc/paper/6112-deep-learning-without-poor-local-minima.pdf.

[25] S. KURMANN, *Some remarks on equations defining coincident root loci*, J. Algebra, 352 (2012), pp. 223–231.

[26] T. LAURENT AND J. VON BRECHT, *Deep linear networks with arbitrary loss: All local minima are global*, in Proceedings of the 35th International Conference on Machine Learning, Proc. Mach. Learn. Res. 80, J. Dy and A. Krause, Stockholmsmässan, Stockholm, Sweden, eds., 2018, pp. 2902–2907, http://proceedings.mlr.press/v80/laurent18a.html.

[27] H. LEE AND B. STURMFELS, *Duality of multiple root loci*, J. Algebra, 446 (2016), pp. 499–526.

[28] C. LENNERZ AND E. SCHOMER, *Efficient distance computation for quadratic curves and surfaces*, in Proceedings of Geometric Modeling and Processing. Theory and Applications, GMP 2002, 2002, pp. 60–69.

[29] H. LU AND K. KAWAGUCHI, *Depth Creates No Bad Local Minima*, preprint, https://arxiv.org/abs/1702.08580, 2017.

[30] D. MEHTA, T. CHEN, T. TANG, AND J. HAUENSTEIN, *The loss surface of deep linear networks viewed through the algebraic geometry lens*, IEEE Trans. Pattern Anal. Mach. Intell., (2021), pp. 1–1, https://doi.org/10.1109/TPAMI.2021.3071289.

[31] A. M. SAXE, J. L. MCCLELLAND, AND S. GANGULI, *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*, in Proceedings of the 2nd International Conference on Learning

Representations, ICLR 2014, Banff, AB, Canada, 2014, Y. Bengio and Y. LeCun, eds., 2014.

[32] A. Seigal and G. Montúfar, *Mixtures and products in two graphical models*, J. Algebr. Stat., 9 (2018), pp. 1–20, https://doi.org/10.18409/jas.v9i1.90.

[33] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, *The implicit bias of gradient descent on separable data*, J. Mach. Learn. Res., 19 (2018), pp. 1–57.

[34] S. Sullivant, *Algebraic Statistics*, Grad. Stud. Math. 194, American Mathematical Society, Providence, RI, 2018.

[35] M. Trager, K. Kohn, and J. Bruna, *Pure and spurious critical points: A geometric study of linear networks*, in International Conference on Learning Representations, 2020, https://openreview.net/forum?id=rkgOlCVYvB.

[36] F. Williams, M. Trager, D. Panozzo, C. Silva, D. Zorin, and J. Bruna, *Gradient dynamics of shallow univariate ReLU networks*, in Advances in Neural Information Processing Systems, Vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., Curran Associates, Inc., Red Hook, NY, 2019, pp. 8378–8387, https://proceedings.neurips.cc/paper/2019/file/1f6419b1cbe79c71410cb320fc094775-Paper.pdf.

[37] L. Zhang, *Depth Creates No More Spurious Local Minima*, preprint, https://arxiv.org/pdf/1901.09827.pdf, 2019.

[38] Y. Zhou and Y. Liang, *Critical points of linear neural networks: Analytical forms and landscape properties*, in International Conference on Learning Representations, 2018, https://openreview.net/forum?id=SysEexbRb.