High-Dimensional Distribution Generation Through Deep Neural Networks

Dmytro Perekrestenko, Léandre Eberhard, and Helmut Bölcskei

Abstract

We show that every d-dimensional probability distribution of bounded support can be generated through deep ReLU networks out of a 1-dimensional uniform input distribution. What is more, this is possible without incurring a cost—in terms of approximation error measured in Wasserstein-distance—relative to generating the d-dimensional target distribution from d independent random variables. This is enabled by a vast generalization of the space-filling approach discovered in [2]. The construction we propose elicits the importance of network depth in driving the Wasserstein distance between the target distribution and its neural network approximation to zero. Finally, we find that, for histogram target distributions, the number of bits needed to encode the corresponding generative network equals the fundamental limit for encoding probability distributions as dictated by quantization theory.

I. Introduction

Deep neural networks have been employed very successfully as generative models for complex natural data such as images [21], [14] and natural language [4], [26]. Specifically, the idea is to train deep networks so that they realize complex high-dimensional probability distributions by transforming samples taken from simple low-dimensional distributions such as uniform or Gaussian [15], [11], [1].

Generative networks with output dimension higher than the input dimension occur, for instance, in language modelling where deep networks are used to predict the next word in a text sequence. Here, the input layer size is determined by the dimension of the word embedding (typically ~ 100) and the output layer, representing a vector of probabilities for each of the words in the vocabulary, is of the size of the vocabulary (typically $\sim 100k$). Another example where the dimension of the output distribution is mandated to be higher than that of the input distribution is given by explicit [15], [24] and implicit [11], [1] density generative networks.

Notwithstanding the practical success of deep generative networks, a profound theoretical understanding of their representational capabilities is still lacking. First results along these lines appear in [16], where it was shown that generative networks can approximate distributions arising from the composition of Barron functions [3]. It remains unclear, however, which distributions can be obtained in such a manner. More recently, it was established [17] that for every given target distribution (of finite third moment) and source distribution, both defined on $\mathbb{R}^d$, there exists a ReLU network whose gradient pushes forward the source distribution to an arbitrarily accurate—in terms of Wasserstein distance—approximation of the target distribution. The aspect of dimensionality increase was addressed in [2], where it is shown that a uniform univariate source distribution can be transformed, by a ReLU network, into a uniform target distribution of arbitrary dimension through a space-filling approach. Besides, [2] also demonstrates how a univariate Gaussian target distribution can be obtained from a univariate uniform source distribution and vice versa. In a general context, the problem of optimal transport [25] between source and target distributions on spaces of different dimensions was studied in [18].

The approximation of distributions through generative networks is inherently related to function approximation and hence to the expressivity of neural networks. A classical result along those lines is the universal approximation theorem [7], [13], which states that single-hidden-layer neural networks with sigmoidal activation function can approximate continuous functions on compact subsets of $\mathbb{R}^n$ arbitrarily well. More recent developments in this area are concerned with the influence of network depth on attainable approximation quality [23], [8], [10]. A theory establishing the fundamental limits of deep neural network expressivity is provided in [6], [9].

D. Perekrestenko is with Google, Switzerland (email: dperekrestenko@google.com)

L. Eberhard is with Upstart Network Inc., USA (e-mail: leandre.eberhard@upstart.com)

H. Bölcskei is with Chair for Mathematical Information Science, ETH Zurich, Switzerland (e-mail: hboelcskei@ethz.ch)

The aim of the present paper is to develop a universal approximation result for generative neural networks. Specifically, we show that every target distribution supported on a bounded subset of $\mathbb{R}^d$ can be approximated arbitrarily well in terms of Wasserstein distance by pushing forward a 1-dimensional uniform source distribution through a ReLU network. The result is constructive in the sense of explicitly identifying the corresponding generative network. Concretely, we proceed in two steps. Given a target distribution, we find the histogram distribution that best approximates it—for a given histogram resolution—in Wasserstein distance. This histogram distribution is then realized by a ReLU network driven by a uniform univariate input distribution. The construction of this ReLU network exploits a space-filling property, vastly generalizing the one discovered in [2], [19]. The main conceptual insight of the present paper is that generating arbitrary d-dimensional target distributions from a 1-dimensional uniform distribution through a deep ReLU network does not come at a cost—in terms of approximation error measured in Wasserstein distance—relative to generating the target distribution from d independent random variables through, e.g., (for arbitrary d) the normalizing flows method [22] and (for $d = 2$) the Box-Muller method [5]. We emphasize that the generating network has to be deep, in fact its depth has to go to infinity to obtain the same Wasserstein-distance error as a construction from d independent random variables would yield. Finally, we find that, for histogram target distributions, the number of bits needed to encode the corresponding generative network equals the fundamental limit for encoding probability distributions as dictated by quantization theory [12].

II. DEFINITIONS AND NOTATION

We start by introducing general notation.

log stands for the logarithm to base 2. For $n_1, n_2 \in \mathbb{N}$, the set of integers in the range $[n_1, n_2]$ is designated as $[n_1 : n_2]$. For $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, we denote the vector obtained by retaining the first t, $t \leq d$, entries by $\mathbf{x}_{[1:t]} := (x_1, x_2, \ldots, x_t) \in \mathbb{R}^t$. $U(\Delta)$ stands for the uniform distribution on the interval Δ; when $\Delta = [0, 1]$, we simply write U. Given a probability density function (pdf) p, we denote its push-forward under the function f as $f\#p$. δ_x refers to the Dirac delta distribution. $\mathfrak{B}^d$ stands for the Borel σ-algebra on $\mathbb{R}^d$, i.e., the smallest σ-algebra on $\mathbb{R}^d$ that contains all open subsets of $\mathbb{R}^d$. For a vector $\mathbf{b} \in \mathbb{R}^d$, we let $\|\mathbf{b}\|_\infty := \max_i |b_i|$, similarly we write $\|\mathbf{A}\|_\infty := \max_{i,j} |\mathbf{A}_{i,j}|$ for the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$. The Cartesian product of the intervals $\mathcal{I}_i$, $i \in [1 : d]$, is denoted by $\times_{i=1}^d \mathcal{I}_i = \mathcal{I}_1 \times \mathcal{I}_2 \times \cdots \times \mathcal{I}_d$. Finally, $\chi_\mathcal{I}$ stands for the indicator function on the set $\mathcal{I}$.

We proceed to define ReLU neural networks.

Definition II.1. *Let $L \in \mathbb{N}$ and $N_0, N_1, \ldots, N_L \in \mathbb{N}$. A ReLU neural network Φ is a map $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ given by*

$$\Phi = \begin{cases} W_1, & L = 1 \\ W_2 \circ \rho \circ W_1, & L = 2 \\ W_L \circ \rho \circ W_{L-1} \circ \rho \circ \cdots \circ \rho \circ W_1, & L \geq 3, \end{cases}$$

where, for $\ell \in [1 : L]$, $W_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, W_\ell(\mathbf{x}) := \mathbf{A}_\ell \mathbf{x} + \mathbf{b}_\ell$ are the associated affine transformations with (weight) matrices $\mathbf{A}_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and (bias) vectors $\mathbf{b}_\ell \in \mathbb{R}^{N_\ell}$, and the ReLU activation function $\rho : \mathbb{R} \to \mathbb{R}$, $\rho(x) := \max(x, 0)$ acts component-wise, i.e., $\rho(x_1, \ldots, x_N) := (\rho(x_1), \ldots, \rho(x_N))$. We denote by $\mathcal{N}_{d,d'}$ the set of all ReLU networks with input dimension $N_0 = d$ and output dimension $N_L = d'$. Moreover, we define the following quantities related to the notion of size of the network Φ:

- *the connectivity $\mathcal{M}(\Phi)$ is the total number of nonzero weights, i.e., entries in the matrices $\mathbf{A}_\ell$, $\ell \in [1 : L]$, and the vectors $\mathbf{b}_\ell$, $\ell \in [1 : L]$*
- *depth $\mathcal{L}(\Phi) := L$*
- *width $\mathcal{W}(\Phi) := \max_{\ell=0,\ldots,L} N_\ell$*

The distance between probability measures will be quantified through Wasserstein distance defined as follows.

Definition II.2. *Let μ and ν be probability measures on $(\mathbb{R}^d, \mathfrak{B}^d)$. A coupling between μ and ν is defined as a probability measure π on $(\mathbb{R}^d \times \mathbb{R}^d, \mathfrak{B}^{2d})$ such that $\pi(A_1 \times \mathbb{R}^d) = \mu(A_1)$ and $\pi(\mathbb{R}^d \times A_2) = \nu(A_2)$, for all $A_1, A_2 \in \mathfrak{B}^d$. Let $\Pi(\mu, \nu)$ be the set of all couplings between μ and ν. The Wasserstein distance between μ and ν is defined as*

$$W(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\| d\pi(\mathbf{x}, \mathbf{y}),$$

where $\|\cdot\|$ denotes Euclidean norm.

We will frequently use the concept of histogram distributions formalized as follows.

Definition II.3. *A random variable X is said to have a general histogram distribution of resolution n on $[0,1]$, denoted as $X \sim \mathcal{G}[0,1]_n^1$, if for some $t_0, t_1, \ldots, t_n \in \mathbb{R}$ such that $0 = t_0 \leq t_1 \leq \cdots \leq t_n = 1$ and if $t_i = t_{i+1}$, then $t_{i+2} \neq t_i$, $\forall i \in [0:(n-2)]$, its pdf is given by*

$$p(x) = \sum_{k=0}^{n-1} w_k \kappa_{[t_k, t_{k+1}]}(x), \ \ with \ \sum_{k=0}^{n-1} w_k d(t_k, t_{k+1}) = 1, \tag{1}$$

and $w_k > 0$ for all $k \in [0:(n-1)]$. Here,

$$\kappa_{[t_k, t_{k+1}]}(x) = \begin{cases} \chi_{[t_k, t_{k+1}]}(x), & if \ t_k < t_{k+1} \\ \delta_{x - t_k}, & if \ t_k = t_{k+1} \end{cases},$$

and

$$d(t_k, t_{k+1}) = \begin{cases} t_{k+1} - t_k, & if \ t_k < t_{k+1} \\ 1, & if \ t_k = t_{k+1} \end{cases}. \tag{2}$$

General histogram distributions allow for bins of arbitrary size and for point singularities, see the right-hand plot in Figure 4. We will, however, mostly be concerned with histogram distributions of uniform tile size, defined as follows.

Definition II.4. *A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)^\top$ is said to have a histogram distribution of resolution n on the d-dimensional unit cube, denoted as $\mathbf{X} \sim \mathcal{E}[0,1]_n^d$, if its pdf is given by*

$$p_{\mathbf{X}}(\mathbf{x}) = \sum_{\mathbf{k}} w_{\mathbf{k}} \chi_{c_{\mathbf{k}}}(\mathbf{x}), \ \ with \ \sum_{\mathbf{k}} w_{\mathbf{k}} = n^d,$$

and $w_{\mathbf{k}} > 0$ for all vectors $\mathbf{k} = (i_1, i_2, \ldots, i_d) \in [0:(n-1)]^d$, referred to as index vectors, and $c_{\mathbf{k}} = [i_1/n, (i_1 + 1)/n] \times [i_2/n, (i_2 + 1)/n] \times \cdots \times [i_d/n, (i_d + 1)/n]$.

Example histogram distributions in the 1- and 2-dimensional case, respectively, are depicted in Figs. 1 and 2. Throughout the paper, to indicate that the random vector $\mathbf{X}$ with distribution $p_{\mathbf{X}}(\mathbf{x})$ satisfies $\mathbf{X} \sim \mathcal{E}[0,1]_n^d$, we shall frequently also write $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]_n^d$.

Remark II.5. *For ease of exposition, in Definitions II.3 and II.4, we let the intervals $[t_k, t_{k+1}]$ and the cubes $c_{\mathbf{k}}$, respectively, be closed, thus allowing the breakpoints to belong to different intervals/cubes. While this comes without loss of generality, for concreteness, it is understood that the value of the pdf at a breakpoint is given by the average across the intervals/cubes containing the breakpoint.*

III. SAWTOOTH FUNCTIONS

As mentioned above, our universal generative network construction is based on a new space-filling property of ReLU networks, vastly generalizing the one discovered in [2], [19]. At the heart of this idea are higher-order sawtooth functions obtained as follows. Consider the sawtooth function $g : \mathbb{R} \to [0,1]$,

$$g(x) = \begin{cases} 2x, & if \ x \in [0, 1/2), \\ 2(1 - x), & if \ x \in [1/2, 1], \\ 0, & else, \end{cases}$$

let $g_1(x) = g(x)$, and define the sawtooth function of order s as the s-fold composition of g with itself according to

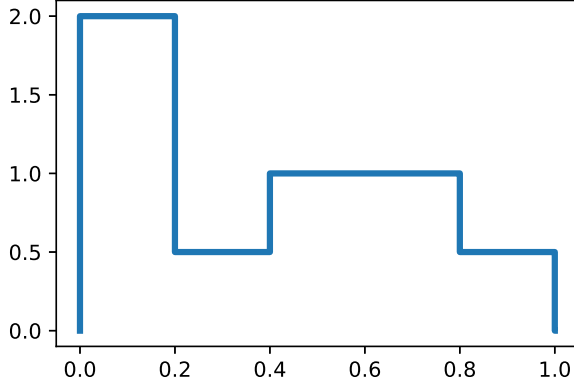$$g_s := \underbrace{g \circ g \circ \cdots \circ g}_{s}, \ \ s \geq 2.$$

Fig. 1: Histogram distribution $\mathcal{E}[0,1]^{1}_{5}$
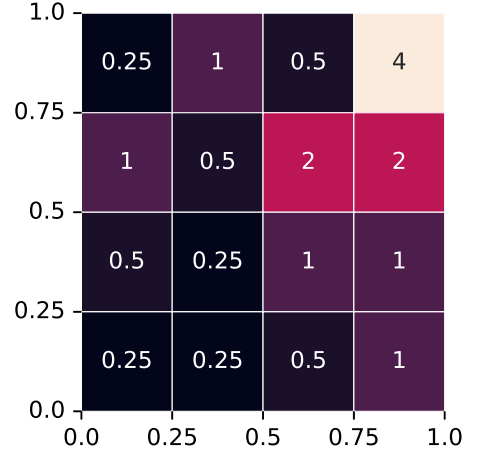

Fig. 2: Histogram distribution $\mathcal{E}[0,1]^{2}_{4}$

Figure 3 depicts the sawtooth functions of orders $1, 2,$ and 3. Next, we note that g can be realized by a 2-layer ReLU network $\Phi_g \in \mathcal{N}_{1,1}$ of connectivity $\mathcal{M}(\Phi_g) = 8$ and depth $\mathcal{L}(\Phi_g) = 2$ according to $\Phi_g = W_2 \circ \rho \circ W_1 = g$ with

$$W_1(x) = \begin{pmatrix} 2 \\ 4 \\ 2 \end{pmatrix} x - \begin{pmatrix} 0 \\ 2 \\ 2 \end{pmatrix}, \quad W_2(\mathbf{x}) = \begin{pmatrix} 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

The sth-order sawtooth function g_s can hence be realized by a ReLU network $\Phi_g^s \in \mathcal{N}_{1,1}$ of connectivity $\mathcal{M}(\Phi_g^s) = 11s - 3$ and depth $\mathcal{L}(\Phi_g^s) = s + 1$ according to $\Phi_g^s = W_2 \circ \rho \circ \underbrace{W_g \circ \rho \circ \cdots \circ W_g \circ \rho}_{s-1} \circ W_1 = g_s$ with

$$W_g(\mathbf{x}) = \begin{pmatrix} 2 & -2 & 2 \\ 4 & -4 & 4 \\ 2 & -2 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - \begin{pmatrix} 0 \\ 2 \\ 2 \end{pmatrix}.$$

We close this section with an important technical ingredient of the generalized space-filling idea presented in Section V.

Lemma III.1. *Let $f(x)$ be a continuous function on $[0,1]$, with $f(0) = 0$. Then, for all $s \in \mathbb{N}$,*

$$f(g_s(x)) = \sum_{k=0}^{2^{s-1}-1} f\big(g(2^{s-1}x - k)\big), \tag{3}$$

and for all $k \in [0 : (2^{s-1} - 1)]$,

$$\mathrm{supp}\big(f\big(g(2^{s-1}x - k)\big)\big) = \left(\frac{k}{2^{s-1}}, \frac{k+1}{2^{s-1}}\right). \tag{4}$$

Proof. We first note that the sawtooth functions $g_s(x)$ satisfy [23]
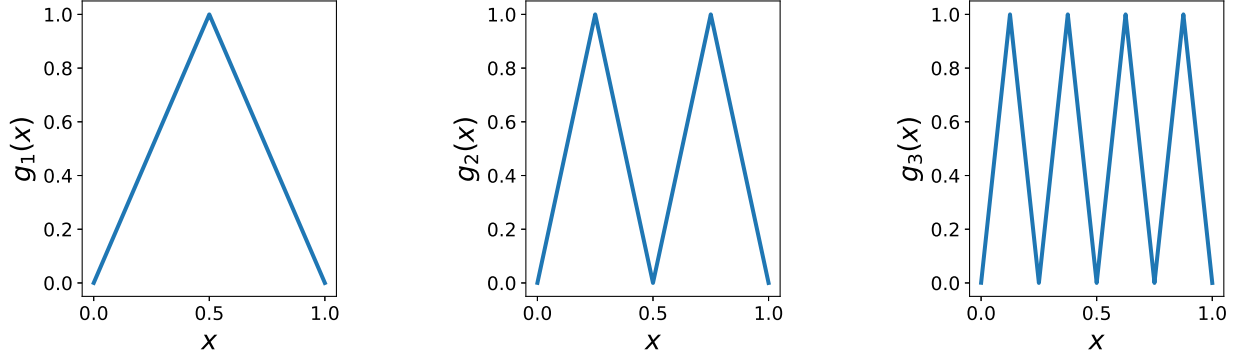
$$g_s(x) = \sum_{k=0}^{2^{s-1}-1} g(2^{s-1}x - k),$$

4

Fig. 3: Sawtooth functions

with $g(2^{s-1}x - k)$ supported on $\left(\frac{k}{2^{s-1}}, \frac{k+1}{2^{s-1}}\right)$. As $f(0) = 0$, the support of $f(g(2^{s-1}x - k))$ coincides with that of $g(2^{s-1}x - k)$, which in turn yields (4). To establish (3), we note that the supports of $g(2^{s-1}x - k)$ are pairwise disjoint across k and hence

$$f(g_s(x)) = f\left(\sum_{k=0}^{2^{s-1}-1} g(2^{s-1}x - k)\right) = \sum_{k=0}^{2^{s-1}-1} f\left(g(2^{s-1}x - k)\right). \qquad \square$$

IV. ReLU NETWORKS GENERATE HISTOGRAM DISTRIBUTIONS

This section establishes a systematic connection between ReLU networks and histogram distributions. Specifically, we show that the pushforward of a uniform distribution under a piecewise linear function results in a histogram distribution. We also identify, for a given histogram distribution, the piecewise linear function generating it under pushforward of a uniform distribution. Combined with the insight that ReLU networks realize piecewise linear functions, the desired connection is established.

We start with an auxiliary result.

Lemma IV.1. *Let $a, b \in \mathbb{R}, a < b, \Delta = [a, b]$, and let $h(x) = mx + s$, for $x \in \mathbb{R}$, with $m \in \mathbb{R}, s \in \mathbb{R}$. Then, $Q = h\#U(\Delta)$ is uniformly distributed on $[ma + s, mb + s]$, for $m > 0$, and on $[mb + s, ma + s]$, for $m < 0$. For $m = 0$, the pdf of Q is given by $\delta_{\cdot - s}$.*

Proof. We start with the case $m \in \mathbb{R} \setminus \{0\}$. The pdf of the pushforward of a random variable with pdf $p(x)$ under the function $f(x)$ is given by

$$q(y) = p(f^{-1}(y)) \left| \frac{d}{dy} f^{-1}(y) \right|.$$

Particularized to $f^{-1}(y) = h^{-1}(y) = \frac{y-s}{m}$ and $p(x) = \frac{1}{b-a}\chi_\Delta(x)$, this yields

$$q(y) = \begin{cases} \frac{1}{m(b-a)}, & \text{if } y \in [ma + s, mb + s] \\ 0, & \text{otherwise,} \end{cases}$$

for $m > 0$, and

$$q(y) = \begin{cases} \frac{1}{m(a-b)}, & \text{if } y \in [mb + s, ma + s] \\ 0, & \text{otherwise,} \end{cases}$$

for $m < 0$. Finally, if $m = 0$, then the entire interval $[a, b]$ is mapped to the point $y = s$ and the corresponding pdf is given by $q(y) = \delta_{y-s}$. $\qquad \square$

We next show that the pushforward of a uniform distribution under a piecewise linear function results in a (general) histogram distribution.

5

Theorem IV.2. *For every piecewise linear continuous function $f : \mathbb{R} \to \mathbb{R}$, such that $f(x) \in [0,1], \forall x \in [0,1]$, and $f(0) = 0, f(1) = 1$, there exists an n so that $f\#U \sim \mathcal{G}[0,1]_n^1$.*

Proof. We split the domain of f into $t \in \mathbb{N}$ pairwise disjoint intervals $I_i = [a_i, b_i], i \in [0:(t-1)]$, each of which f is linear on, specifically $f(x) = m_i x + s_i$, $x \in I_i$. Using the law of total probability and the chain rule, the pdf of $q = f\#U$ can accordingly be represented as

$$q(y) = \sum_{j=0}^{t-1} q(y|u \in I_j)\mathbb{P}(u \in I_j). \tag{5}$$

As U is uniform, it is also uniform conditional on being in a given interval I_j. By Lemma IV.1 it therefore follows that $q(y|u \in I_j)$ can be written as

$$q(y|u \in I_j) = \begin{cases} \frac{\chi_{R_j}(y)}{|R_j|}, & \text{if } m_j \neq 0, \\ \delta_{y-s_j}, & \text{if } m_j = 0, \end{cases}$$

where $R_j = [m_j a_j + s_j, m_j b_j + s_j]$ if $m_j > 0$, and $R_j = [m_j b_j + s_j, m_j a_j + s_j]$ if $m_j < 0$. Noting that by continuity of f and the boundary conditions $f(0) = 0, f(1) = 1$, we have $\bigcup_j R_j = [0,1]$, it follows that $q(y)$ in (5) corresponds to a general histogram distribution according to (1) with $n = t$ and

$$w_j = \begin{cases} \frac{\mathbb{P}(u \in I_j)}{|R_j|}, & \text{if } m_j \neq 0, \\ \mathbb{P}(u \in I_j), & \text{if } m_j = 0. \end{cases} \qquad \square$$

We will also need the converse to the result just established, in particular a constructive version thereof explicitly identifying the piecewise linear function that leads to a given general histogram distribution under pushforward of a uniform distribution on the interval $[0,1]$.

Theorem IV.3. *Let $p(x)$ be the pdf of $X \sim \mathcal{G}[0,1]_n^1$ with weights w_k, $k \in [0:(n-1)]$, and breakpoints $0 = t_0 \leq t_1 \leq \cdots \leq t_n = 1$, and set $b_0 = 0$, $b_i = \sum_{k=0}^{i-1} w_k d(t_k, t_{k+1})$, $i \in [1:n]$, where*

$$d(t_k, t_{k+1}) = \begin{cases} t_{k+1} - t_k, & \text{if } t_k < t_{k+1} \\ 1, & \text{if } t_k = t_{k+1} \end{cases}.$$

Further, define a_i, $i \in [0:(n-1)]$, as follows: If $t_0 = t_1$, then $a_0 = 0$ and $a_1 = \frac{1}{w_1}$. If $t_0 \neq t_1$, then $a_0 = \frac{1}{w_0}$. For $k \in [1:(n-2)]$, if $t_k = t_{k+1}$, then $a_k = -\frac{1}{w_{k-1}}$ and $a_{k+1} = \frac{1}{w_{k+1}}$, and, if $t_{k-1} \neq t_k \neq t_{k+1}$, then $a_k = \frac{1}{w_k} - \frac{1}{w_{k-1}}$. Finally, if $t_{n-1} \neq 1$, then $a_{n-1} = \frac{1}{w_{n-1}} - \frac{1}{w_{n-2}}$. Then,

$$f(x) = \sum_{i=0}^{n-1} a_i \rho(x - b_i) \tag{6}$$

is the piecewise linear function satisfying $f\#U = p$.

Proof. Let $I_i := [b_i, b_{i+1}]$, $i \in [0:(n-1)]$. Then, $\bigcup_{i \in [0:(n-1)]} I_i = [0,1]$ and, for all $i \in [0:(n-1)]$, the function $f(x)$ in (6) is linear on I_i with slope given by

$$\sum_{j=0}^{i} a_j = \begin{cases} 1/w_i, & \text{if } t_i \neq t_{i+1} \\ 0, & \text{if } t_i = t_{i+1} \end{cases}.$$

Next, note that under $f(x)$ the interval I_i is mapped to the interval $[t_i, t_{i+1}]$ if $t_i \neq t_{i+1}$ and to the singleton $\{t_i\}$ if $t_i = t_{i+1}$. The proof is finalized upon using $\mathbb{P}(u \in I_i) = b_{i+1} - b_i$ to conclude that (cf. the proof of Theorem IV.2) $\mathbb{P}(u \in I_i)/|(1/w_i)(b_{i+1} - b_i)| = w_i$, for $t_i \neq t_{i+1}$, and $\mathbb{P}(u \in I_i) = b_{i+1} - b_i = w_i d(t_i, t_{i+1}) = w_i$ in the case $t_i = t_{i+1}$. $\square$

An example of a piecewise linear function and the corresponding general histogram distribution according to Theorems IV.2 and IV.3 is provided in Figure 4. Theorems IV.2 and IV.3 are of independent interest as they allow to conclude that ReLU networks, by virtue of always realizing piecewise linear functions, produce general histogram
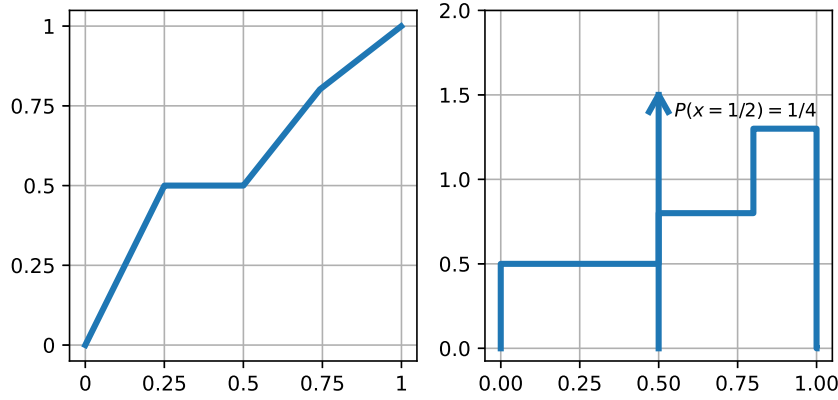
Fig. 4: A piecewise linear function f (left) and the corresponding general histogram distribution $f \# U$ (right).

distributions when pushing forward uniform distributions. In the remainder of the paper, we shall, however, work with histogram distributions $\mathcal{E}[0,1]_n^d$ only, in particular for $d = 1$, in which case Theorem IV.3 takes on a simpler form spelled out next.

Corollary IV.4. *Let $p(x)$ be the pdf of $X \sim \mathcal{E}[0,1]_n^1$ with weights w_k, $k \in [0{:}n]$, and let $a_0 = \frac{1}{w_0}$, $a_i = \frac{1}{w_i} - \frac{1}{w_{i-1}}$, $i \in [1{:}(n-1)]$, $b_0 = 0$, $b_i = \frac{1}{n} \sum_{k=0}^{i-1} w_k$, $i \in [1{:}n]$. Then, $p = f \# U$ with the piecewise linear function*

$$f(x) = \sum_{i=0}^{n-1} a_i \rho(x - b_i).$$

We shall often need the explicit form of $f(x)$ in Corollary IV.4 on its intervals of linearity $[b_\ell, b_{\ell+1}]$. A direct calculation reveals that

$$f(x) = \frac{x}{w_\ell} - \frac{\sum_{i=0}^{\ell-1} w_i}{n w_\ell} + \frac{\ell}{n}, \quad x \in [b_\ell, b_{\ell+1}].$$

V. INCREASING DISTRIBUTION DIMENSIONALITY

This section is devoted to the aspect of distribution dimensionality increase through deep ReLU networks. Specifically, for a given random vector $\mathbf{X} \sim \mathcal{E}[0,1]_n^d$ with (histogram) distribution $p_{\mathbf{X}}(\mathbf{x})$ of resolution n, we construct a piecewise linear map $M : [0,1] \to [0,1]^d$ such that the pushforward $M \# U$ approximates $p_{\mathbf{X}}(\mathbf{x})$ arbitrarily well. The main ingredient of our construction is a vast generalization of the space-filling property of sawtooth functions discovered in [19]. Informally speaking, the novel space-filling property we describe allows to completely fill the d-dimensional target space according to a prescribed target histogram distribution by transporting probability mass from the 1-dimensional uniform distribution U to d-dimensional space.

We first develop some intuition behind this construction. Specifically, we consider the 2-dimensional case and visualize the idea of approximating a 2-dimensional target distribution through pushforward of U by the sawtooth functions $g_s(x)$ (depicted in Figure 3) as painting the curve $g_s(x)$ with probability mass taken from U. The geodesic distance traveled by the brush distributing probability mass along $g_s(x)$ goes to infinity according to 2^s as $s \to \infty$. This follows by noting that the number of teeth is 2^s and for $s \to \infty$, the length of the individual teeth (in fact their halves) approaches 1. Therefore, as $s \to \infty$ the square $[0,1]^2$ will be filled with paint completely. Moreover, as x traverses from 0 to 1, the speed at which probability mass is allocated to the marginal dimensions, i.e., along the x_1-and x_2-axes, is constant. To see this, simply note that along the x_2-axis the speed of the brush is given by the derivative of $g_s(x)$, which by virtue of $g_s(x)$ consisting of piecewise linear segments, is constant. Likewise, as the inverse of a linear function is again a linear function, the brush moves with constant speed along the x_1-axis as well. This guarantees that the resulting 2-dimensional probability distribution along with its marginals and conditional

7

distributions are all uniform. The rate at which the joint distribution approaches a 2-dimensional uniform distribution can be quantified in terms of Wasserstein distance by defining the transport map $M : x \to (x, g_s(x))$ and noting that $W(M\#U, U([0,1]^2)) \leq \frac{\sqrt{2}}{2^s}$ [2]. What is noteworthy here is that the map M takes probability mass from $\mathbb{R}$ to $\mathbb{R}^2$ in a space-filling fashion, i.e., we get a dimensionality increase as $s \to \infty$.

By adjusting the "paint plan", this idea can now be generalized to 2-dimensional histogram target distributions that are constant with respect to one of the dimensions, here, for concreteness, the x_1-dimension. Specifically, we replace $g_s(x)$ in the construction above by $f(g_s(x))$, where the piecewise linear function $f(x)$ determines the paint plan resulting in the desired weights (across the x_2-axis) according to Corollary IV.4. We refer to Figure 5 for an illustration of the idea. While the outer function $f(x)$ determines how much time the paint brush spends in a given interval along the x_2-axis, the inner function $g_s(x)$ takes care of filling the unit square as $s \to \infty$. The larger the slope of $f(x)$ on a given interval along the x_2-axis, the less time the brush spends in that interval and the smaller the amount of probability mass allocated to the interval. Concretely, by Corollary IV.4 the amount of probability mass deposited in a given interval is inversely proportional to the slope of $f(x)$ on that interval.

Finally, consider the 2-dimensional histogram distribution $p_{X_1,X_2}(x_1, x_2) \in \mathcal{E}[0,1]_n^2$ and note that it can be represented as follows

$$
\begin{aligned}
p_{X_1,X_2}(x_1, x_2) &= \sum_{k_1,k_2} w_{k_1,k_2} \chi_{c_{k_1,k_2}}(x_1, x_2) = \sum_{k_1,k_2} w_{k_1} w_{k_2|k_1} \chi_{c_{k_1}}(x_1) \chi_{c_{k_2}}(x_2) \\
&= \sum_{k_1} w_{k_1} \chi_{c_{k_1}}(x_1) \sum_{k_2} w_{k_2|k_1} \chi_{c_{k_2}}(x_2) \\
&= \sum_{i=0}^{n-1} p_{X_1}\big(x_1 \in [i/n, (i+1)/n]\big)\, p_{X_2|X_1}\big(x_2 | x_1 \in [i/n, (i+1)/n]\big),
\end{aligned}
\tag{7}
$$

where $w_{k_1} = \frac{1}{n} \sum_{k_2} w_{k_1,k_2}$, $w_{k_2|k_1} = w_{k_1,k_2}/w_{k_1}$, $p_{X_1}(x_1 \in [i/n, (i+1)/n]) = w_i \chi_{c_i}(x_1)$ denotes the restriction of the marginal histogram distribution p_{X_1} (see Lemma V.1 below) to the interval $[i/n, (i+1)/n]$, and $p_{X_2|X_1}\big(x_2 | x_1 \in [i/n, (i+1)/n]\big) = \sum_{k_2} w_{k_2|i} \chi_{c_{k_2}}(x_2)$, for each $i \in [0 : (n-1)]$, can be viewed as a 2-dimensional histogram distribution that is constant with respect to x_1, and which we assume to be generated by $f_{X_2}^{(i)}(g_s(x))$ according to the procedure described in the previous paragraph.

Now, in order to "paint" the general 2-dimensional histogram distribution in (7), we have to "squeeze" the space-filling curves $f_{X_2}^{(i)}(g_s(x))$ into the respective boxes $[i/n, (i+1)/n] \times [0,1]$. This is effected by exploiting that $g_s(x)$ is compactly supported on $[0,1]$ for all $s \in \mathbb{N}$, which allows us to realize the desired localization according to $f_{X_2}^{(i)}(g_s(nx - i))$. The resulting localized space-filling curves are then stitched together by adding them up according to $\sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nx - i))$. Denoting the piecewise linear function that generates the marginal histogram distribution p_{X_1} according to Corollary IV.4 as[1] $f_{X_1}^{\mathbf{z}_1}$, i.e., $f_{X_1}^{\mathbf{z}_1}\#U = p_{X_1}$, the transport map $M : x \to \left(f_{X_1}^{\mathbf{z}_1}(x), \sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(n f_{X_1}^{\mathbf{z}_1}(x) - i)) \right)$ when applied to U generates p_{X_1,X_2} in (7) asymptotically in s. To see this, we first note that the second component of M pushes forward, by $\sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nx - i))$, the random variable $f_{X_1}^{\mathbf{z}_1}\#U$ resulting from the pushforward of U by the first component of M. This allows us to read off the conditional distributions $p_{X_2|X_1}$. Specifically, thanks to the individual components in $\sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nx - i))$ being disjointly supppported on $[i/n, (i+1)/n]$, we can conclude that the distribution of the 2-dimensional random variable $M\#U$ satisfies $p_{X_2|X_1}\big(x_2 | x_1 \in [i/n, (i+1)/n]\big) = \sum_{k_2} w_{k_2|i} \chi_{c_{k_2}}(x_2), i \in [0 : (n-1)]$, as desired. Next, noting that the distribution p_{X_1} of $f_{X_1}^{\mathbf{z}_1}\#U$ has components $p_{X_1}\big(x_1 \in [i/n, (i+1)/n]\big) = w_i \chi_{c_i}(x_1)$ supported disjointly on $[i/n, (i+1)/n]$, it follows from $p_{X_1,X_2}(x_1, x_2) = p_{X_1}(x_1)\, p_{X_2|X_1}(x_2|x_1)$ that the distribution of $M\#U$ is given by

$$
p_{X_1,X_2}(x_1, x_2) = \sum_{i=0}^{n-1} p_{X_1}\big(x_1 \in [i/n, (i+1)/n]\big)\, p_{X_2|X_1}\big(x_2 | x_1 \in [i/n, (i+1)/n]\big),
$$

which is (7). An example illustrating the overall construction can be found in Figure 6.

We are now ready to formalize the idea just described and generalize it to target distributions of arbitrary dimension. To this end, we start with a technical lemma stating that all marginal and conditional distributions

---

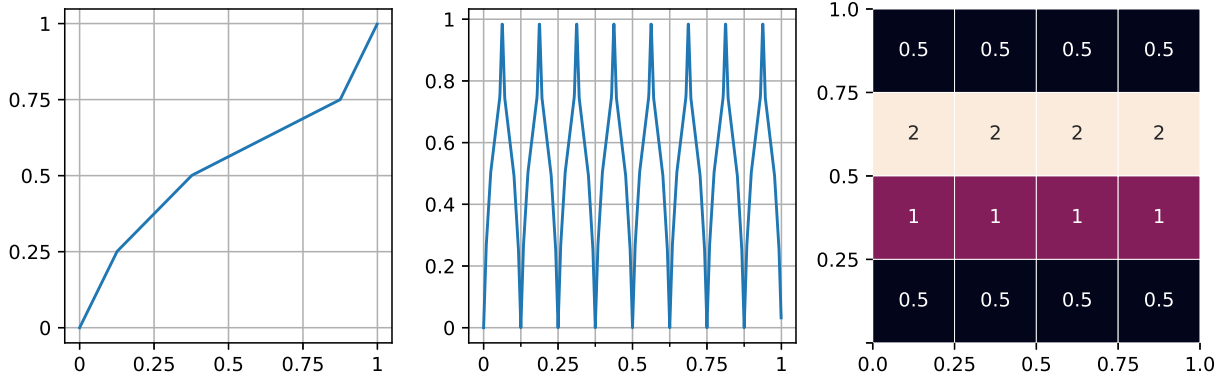[1]The choice of the superscript $\mathbf{z}_1$ in $f_{X_1}^{\mathbf{z}_1}$ will become clear in Definition V.2 below.

8

Fig. 5: Generating a histogram distribution via the transport map $x \to (x, f(g_s(x)))$. Left—the function $f(x)$, center—$f(g_4(x))$, right—a heatmap of the resulting histogram distribution.

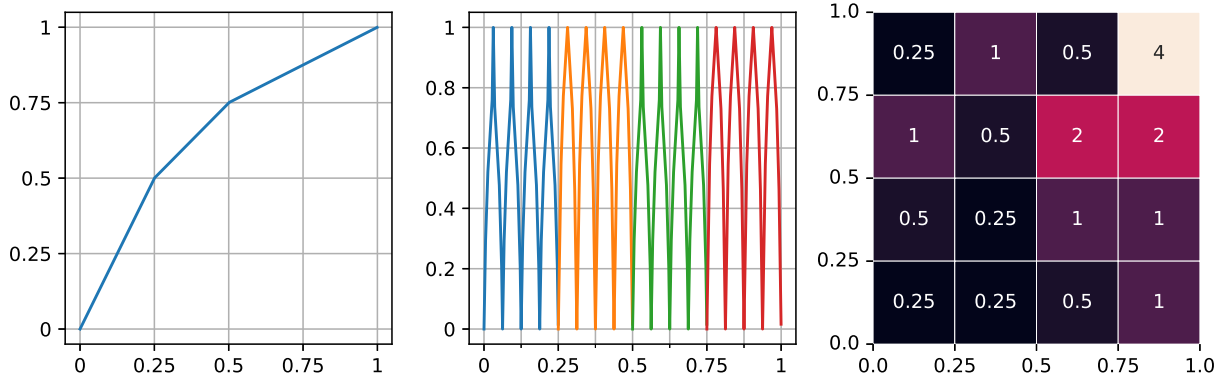Fig. 6: Generating a 2-D histogram distribution via the transport map $x \to (f_{X_1}^{\mathbf{z}_1}(x), \sum_{i=0}^3 f_{X_2}^{(i)}(g_3(4f_{X_1}^{\mathbf{z}_1}(x) - i)))$. Left—the function $f_{X_2}^{(1)} = f_{X_2}^{(3)} = f_{X_1}^{\mathbf{z}_1}$, center—$\sum_{i=0}^3 f_{X_2}^{(i)}(g_3(4f_{X_1}^{\mathbf{z}_1}(x) - i))$, right—a heatmap of the resulting histogram distribution. We took $f_{X_2}^{(0)} = f_{X_2}^{(2)}$ to be given by the function depicted on the left in Figure 5.

of a d-dimensional histogram distribution are themselves histogram distributions, a result that was already used implicitly in the description of our main idea in the 2-dimensional case above.

Lemma V.1. *Let $p_{\mathbf{X}}(\mathbf{x})$ be the pdf of the random vector $\mathbf{X} \sim \mathcal{E}[0,1]_n^d$. Then, for all $t \in [1:(d-1)]$, its marginal distributions satisfy $p_{X_1,\ldots,X_t}(\mathbf{x}_{[1:t]}) \in \mathcal{E}[0,1]_n^t$. Moreover, for all $t \in [1:(d-1)]$ and $\mathbf{z} = (z_1, z_2, \ldots, z_t) \in [0:(n-1)]^t$, defining $c_{\mathbf{z}} = [z_1/n, (z_1+1)/n] \times [z_2/n, (z_2+1)/n] \times \cdots \times [z_t/n, (z_t+1)/n]$, the conditional distributions $p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]} \in c_{\mathbf{z}})$ are independent of the specific value of $\mathbf{x}_{[1:t]} \in c_{\mathbf{z}}$ and obey $p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]} \in c_{\mathbf{z}}) \in \mathcal{E}[0,1]_n^1$.*

Proof. The proof of the first statement follows by noting that, for all $t \in [1:(d-1)]$,

$$
\begin{aligned}
p_{X_1,\ldots,X_t}(\mathbf{x}_{[1:t]}) &= \int_{[0,1]^{d-t}} \sum_{\mathbf{k}} w_{\mathbf{k}} \chi_{c_{\mathbf{k}}}(\mathbf{x}) dx_{t+1} dx_{t+2} \ldots dx_d \\
&= \sum_{\mathbf{z}} w_{\mathbf{z}} \chi_{c_{\mathbf{z}}}(\mathbf{x}_{[1:t]}),
\end{aligned}
\tag{8}
$$

where $\mathbf{z} = \mathbf{k}_{[1:t]} = (z_1, z_2, \ldots, z_t)$ with $\mathbf{k}$ according to Definition II.4, and $w_{\mathbf{z}} = (1/n)^{d-t} \sum_{i_{t+1},\ldots,i_d} w_{\mathbf{k}} > 0$. With $\sum_{\mathbf{k}} w_{\mathbf{k}} = n^d$ from Definition II.4, we get $\sum_{\mathbf{z}} w_{\mathbf{z}} = n^{t-d} \sum_{\mathbf{k}} w_{\mathbf{k}} = n^t$, which establishes that (8) constitutes

9

a valid histogram distribution in $\mathcal{E}[0,1]_n^t$.

To prove the second statement, we first note that for all $t \in [1\!:\!(d-1)]$,

$$p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]}) = \frac{p_{X_1,\ldots,X_{t+1}}(\mathbf{x}_{[1:(t+1)]})}{p_{X_1,\ldots,X_t}(\mathbf{x}_{[1:t]})} = \frac{\sum_{(\mathbf{z},z_{t+1})} w_{(\mathbf{z},z_{t+1})}\chi_{c_{(\mathbf{z},z_{t+1})}}(\mathbf{x}_{[1:(t+1)]})}{\sum_{\mathbf{z}} w_{\mathbf{z}}\chi_{c_{\mathbf{z}}}(\mathbf{x}_{[1:t]})}.$$

Next, for a given $\mathbf{z}' \in [1\!:\!(n-1)]^t$, we have

$$p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]} \in c_{\mathbf{z}'}) = \frac{\sum_{z_{t+1}} w_{(\mathbf{z}',z_{t+1})}\chi_{c_{(\mathbf{z}',z_{t+1})}}(\mathbf{x}_{[1:(t+1)]})}{w_{\mathbf{z}'}}$$
$$= \sum_{z_{t+1}} \frac{w_{(\mathbf{z}',z_{t+1})}}{w_{\mathbf{z}'}}\chi_{c_{z_{t+1}}}(x_{t+1}),$$

which allows us to conclude that, for all $t \in [1\!:\!(d-1)]$ and $\mathbf{z} = (z_1, z_2, \ldots, z_t) \in [0\!:\!(n-1)]^t$, the conditional distribution $p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]} \in c_{\mathbf{z}})$ is independent of the specific value of $\mathbf{x}_{[1:t]} \in c_{\mathbf{z}}$ and, thanks to $\sum_{z_{t+1}} \frac{w_{(\mathbf{z}',z_{t+1})}}{w_{\mathbf{z}'}} = n$, belongs to the class $\mathcal{E}[0,1]_n^1$. □

As a consequence of Lemma V.1 it follows—from the chain rule—that the joint distribution of a histogram distribution $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]_n^d$ is fully specified by the conditional distributions $p_{X_{t+1}|X_1,\ldots,X_t}(x_{t+1}|\mathbf{x}_{[1:t]} \in c_{\mathbf{z}}), t \in [1\!:\!(d-1)], \mathbf{z} = (z_1, z_2, \ldots, z_t) \in [0\!:\!(n-1)]^t$, and the marginal distribution $p_{X_1}(x_1)$.

We next define the auxiliary functions F_r and Z_r needed in the construction of the d-dimensional generalization of the 2-dimensional transport map $M : x \to \left(f_{X_1}^{\mathbf{z}_1}(x), \sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nf_{X_1}^{\mathbf{z}_1}(x) - i)) \right)$.

Definition V.2. *For $\mathbf{k} = [k_1, \ldots, k_t] \in [0\!:\!(n-1)]^t, t \in \mathbb{N}$, define $c_{\mathbf{k}} = \left[\frac{k_1}{n}, \frac{k_1+1}{n} \right] \times \left[\frac{k_2}{n}, \frac{k_2+1}{n} \right] \times \cdots \times \left[\frac{k_t}{n}, \frac{k_t+1}{n} \right]$. Let $\mathbf{z} = (z_1, z_2, \ldots, z_d) \in [0\!:\!(n-1)]^d$, set $\mathbf{z}_i = \mathbf{z}_{[1:(i-1)]}$, for $i \in [2\!:\!d]$, and fix a histogram distribution $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]_n^d$ specified by $p_{X_i}^{\mathbf{z}_i} = p_{X_i|X_1,\ldots,X_{i-1}}(x_i|\mathbf{x}_{[1:(i-1)]} \in c_{\mathbf{z}_i})$, for $i \in [2\!:\!d]$, and[2] $p_{X_1}^{\mathbf{z}_1}(x_1) = p_{X_1}(x_1)$. For $i \in [1\!:\!d]$, let $f_{X_i}^{\mathbf{z}_i}$ be the piecewise linear function that, according to Corollary IV.4, satisfies $f_{X_i}^{\mathbf{z}_i}\#U = p_{X_i}^{\mathbf{z}_i}$, and define recursively, for all $s \in \mathbb{N}$,*

$$F_r(x, \mathbf{z}_{r+1}, s) := g_s\left(nf_{X_r}^{\mathbf{z}_r}\left(F_{r-1}(x, \mathbf{z}_r, s)\right) - z_r\right), \quad r \in [1\!:\!(d-1)], \tag{9}$$

with the initialization

$$F_0(x, \mathbf{z}_1, s) := x.$$

Further, define the functions Z_r according to

$$Z_r(x, s) := \sum_{\mathbf{z}_r} f_{X_r}^{\mathbf{z}_r}\left(F_{r-1}(x, \mathbf{z}_r, s)\right), \quad r \in [2\!:\!d],$$

and

$$Z_1(x, s) := f_{X_1}^{\mathbf{z}_1}(x).$$

With the quantities just defined, we can write

$$M : x \to \left(f_{X_1}^{\mathbf{z}_1}(x), \sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nf_{X_1}^{\mathbf{z}_1}(x) - i)) \right) = (Z_1(x, s), Z_2(x, s)).$$

In the d-dimensional case, the space-filling transport map that takes a 1-dimensional uniform distribution into a given d-dimensional histogram distribution, or more precisely a sufficiently accurate approximation thereof, will be seen to be given by

$$M : x \to (Z_1(x, s), Z_2(x, s), \ldots, Z_d(x, s)). \tag{10}$$

Theorem V.6 below, the central result of this section, makes this formal. The material from here on up to Theorem V.6 is all preparatory and technical. We recommend that it be skipped at first reading and suggest to proceed to Theorem V.6, in particular the intuition behind the construction of (10) provided right after the proof of Theorem V.6.

---

[2] Formally, $\mathbf{z}_1$, albeit not defined, would correspond to a 0-dimensional quantity. It is used throughout the paper only for notational convenience.

We do recommend, however, to first visit Figure 7, which illustrates the F_r-functions and their role in generating the target histogram distribution.

The following lemma establishes support properties of the F_r-functions and corresponding consequences for the Z_r-functions.

Lemma V.3. *Let $F_i, i \in [0\!:\!(d-1)], \mathbf{z}_i, i \in [2\!:\!d], Z_i, f_{X_i}^{\mathbf{z}_i}, i \in [1\!:\!d]$, be as in Definition V.2. Then, for all $r \in [2\!:\!d]$, we have*

$$\bigcap_{\mathbf{z}_r} \mathrm{supp}\big(F_{r-1}(x, \mathbf{z}_r, s)\big) = \varnothing,$$

and hence for every $r \in [2\!:\!d]$, for all $\mathbf{t}_r \in [0\!:\!(n-1)]^{r-1}$, it holds that

$$f_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(x, \mathbf{t}_r, s)\big) > 0 \implies Z_r(x, s) = f_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(x, \mathbf{t}_r, s)\big), \quad x \in [0, 1].$$

Proof. The proof is through induction across r. We start with the base case $r = 2$. Since the vector $\mathbf{z}_2$ is, in fact, a scalar, we can write $\mathbf{z}_2 = z_1$ and note that

$$F_1(x, \mathbf{z}_2, s) = F_1(x, z_1, s) = g_s\big(nf_{X_1}^{\mathbf{z}_1}(x) - z_1\big)$$

and

$$Z_2(x, s) = \sum_{\mathbf{z}_2} f_{X_2}^{\mathbf{z}_2}\big(g_s\big(nf_{X_1}^{\mathbf{z}_1}(x) - z_1\big)\big) = \sum_{z_1=0}^{n-1} f_{X_2}^{(z_1)}\big(g_s\big(nf_{X_1}^{\mathbf{z}_1}(x) - z_1\big)\big).$$

Fix an arbitrary $\hat{x} \in [0, 1]$. Since $f_{X_1}^{\mathbf{z}_1}(x) \in [0, 1]$, it follows that $f_{X_1}^{\mathbf{z}_1}(\hat{x}) \in [t_1/n, (t_1+1)/n]$ for some $t_1 \in [0\!:\!(n-1)]$. Then, $\big(nf_{X_1}^{\mathbf{z}_1}(\hat{x}) - z_1\big) \in [0, 1]$ if and only if $z_1 = t_1$. Further, as for $x \notin [0, 1]$, $g_s(x) = 0$, we have $F_1(\hat{x}, z_1, s) = g_s\big(nf_{X_1}^{\mathbf{z}_1}(\hat{x}) - z_1\big) = 0$, for all $z_1 \neq t_1$. Combined with the fact that $\hat{x} \in [0, 1]$ was chosen arbitrarily, this implies that for every $x \in [0, 1]$, there exists a $t_1 \in [0\!:\!(n-1)]$ such that for all $z_1 \neq t_1$, it holds that $F_1(x, z_1, s) = 0$. This can equivalently be expressed as

$$\bigcap_{z_1=0}^{n-1} \mathrm{supp}\big(F_1(x, z_1, s)\big) = \varnothing. \tag{11}$$

Next, since $f_{X_2}^{z_1}(0) = 0$, for all $z_1 \in [0\!:\!(n-1)]$, it follows from (11) that, for all $t_1 \in [0\!:\!(n-1)]$,

$$f_{X_2}^{(t_1)}(F_1(x, t_1, s)) > 0 \implies Z_2(x, s) = \sum_{z_1=0}^{n-1} f_{X_2}^{(z_1)}\big(F_1(x, z_1, s)\big) = f_{X_2}^{(t_1)}(F_1(x, t_1, s)), \quad x \in [0, 1].$$

This establishes the base case. To prove the induction step, we assume that the statement holds for some $r \in [2\!:\!(d-1)]$. Then, by the induction assumption,

$$\bigcap_{\mathbf{z}_r} \mathrm{supp}\big(F_{r-1}(x, \mathbf{z}_r, s)\big) = \varnothing,$$

or, equivalently, for every $x \in [0, 1]$, there exists a $\mathbf{t}_r \in [0:(n-1)]^{r-1}$ such that for all $\mathbf{z}_r \neq \mathbf{t}_r$, we have $F_{r-1}(x, \mathbf{z}_r, s) = 0$. Now, fix an arbitrary $\hat{x} \in [0, 1]$ together with its corresponding $\mathbf{t}_r \in [0:(n-1)]^{r-1}$ such that $F_{r-1}(\hat{x}, \mathbf{z}_r, s) = 0$, for all $\mathbf{z}_r \neq \mathbf{t}_r$, and consider

$$F_r(\hat{x}, \mathbf{z}_{r+1}, s) = g_s\big(nf_{X_r}^{\mathbf{z}_r}\big(F_{r-1}(\hat{x}, \mathbf{z}_r, s)\big) - z_r\big) = \begin{cases} 0, & \text{if } \mathbf{z}_r \neq \mathbf{t}_r, \\ g_s\big(nf_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(\hat{x}, \mathbf{t}_r, s)\big) - z_r\big), & \text{if } \mathbf{z}_r = \mathbf{t}_r. \end{cases} \tag{12}$$

Again, as $f_{X_r}^{\mathbf{t}_r}(x) \in [0, 1]$, we can conclude that, for an arbitrarily fixed $\hat{x} \in [0, 1]$, there is a $t_r \in [0\!:\!(n-1)]$ such that $f_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(\hat{x}, \mathbf{t}_r, s)\big) \in [t_r/n, (t_r+1)/n]$. Then, $\big(nf_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(\hat{x}, \mathbf{t}_r, s)\big) - z_r\big) \in [0, 1]$ if and only if $z_r = t_r$. Thanks to $g_s(x) = 0$, for $x \notin [0, 1]$, (12) becomes

$$F_r(\hat{x}, \mathbf{z}_{r+1}, s) = \begin{cases} 0, & \text{if } \mathbf{z}_{r+1} \neq \mathbf{t}_{r+1}, \\ g_s\big(nf_{X_r}^{\mathbf{t}_r}\big(F_{r-1}(\hat{x}, \mathbf{t}_r, s)\big) - t_r\big) = F_r(\hat{x}, \mathbf{t}_{r+1}, s), & \text{if } \mathbf{z}_{r+1} = \mathbf{t}_{r+1}. \end{cases}$$
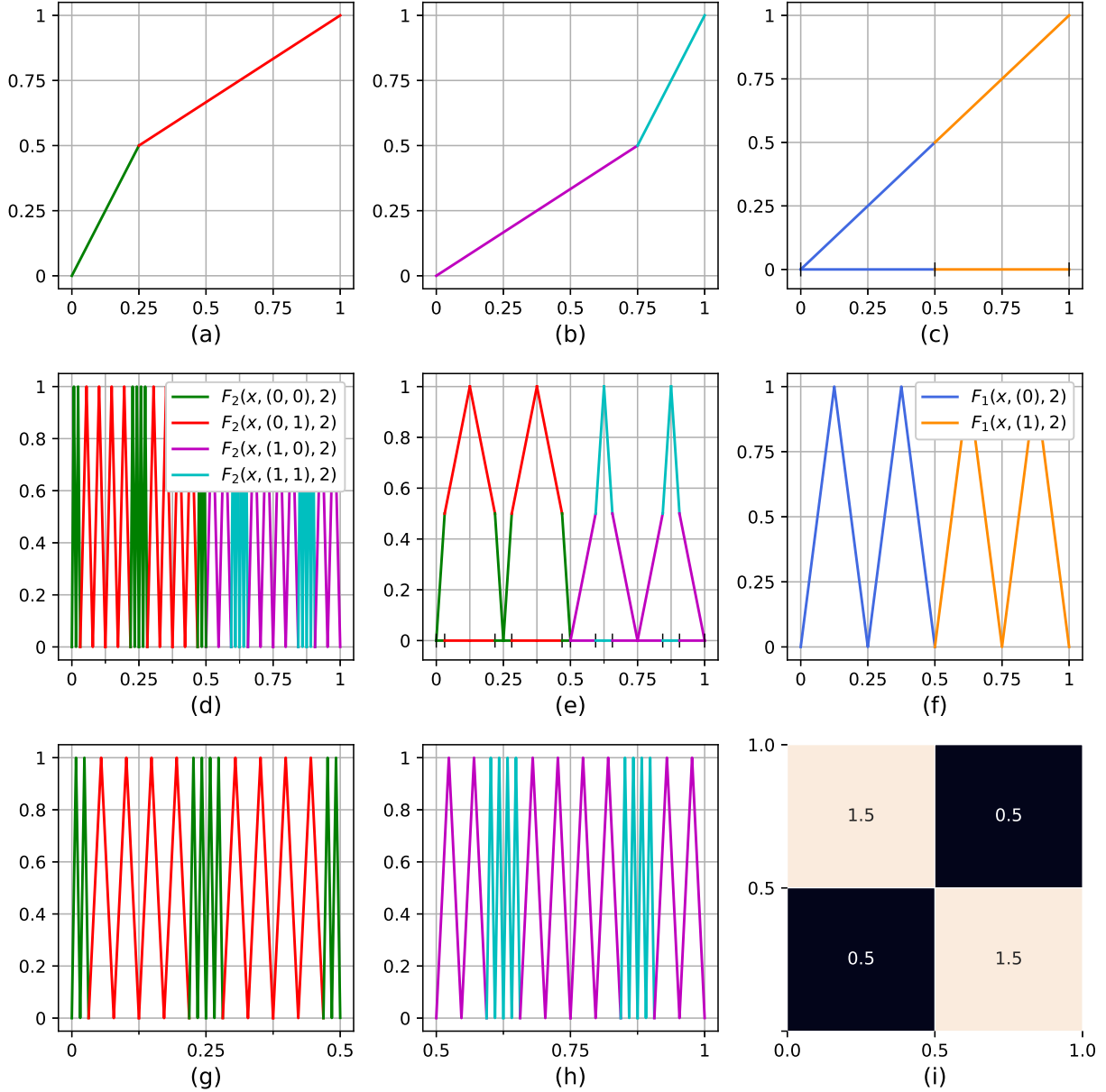
Fig. 7: An example illustrating the functions $F_r(x, \mathbf{z}_{r+1}, s)$ for $r = 1$ and $r = 2$. The corresponding target histogram distribution $p_{X_1,X_2}(x_1, x_2)$ is visualized in subplot (i). The function $Z_1(x, s) = f_{X_1}^{\mathbf{z}_1}(x) = x$ characterizing the marginal $p_{X_1} = U$ is shown in subplot (c). The functions $f_{X_2}^{(0)}(x)$ and $f_{X_2}^{(1)}(x)$ characterizing $p_{X_2}^{(0)}$ and $p_{X_2}^{(1)}$, respectively, are depicted in subplots (a) and (b). Subplot (e) shows $Z_2(x, 2)$ and subplot (f) visualizes $F_1(x, (0), 2)$ and $F_1(x, (1), 2)$. The functions $F_2(x, \mathbf{z}_3, 2)$ are shown in subplot (d). Subplots (g) and (h) depict zoomed-in versions of subplot (d). The functions $F_2(x, \mathbf{z}_3, 2)$ have disjoint support sets, but, in contrast to $F_1(x, (0), 2)$ and $F_1(x, (1), 2)$, their support sets are not connected. The support sets of the colored pieces in subplot (d) match those of the respective colored pieces in subplot (e), likewise for subplots (f) and (c).

12

Combined with the fact that $\hat{x} \in [0,1]$ was chosen arbitrarily, this implies that for every $x \in [0,1]$, there exists a $\mathbf{t}_{r+1} \in [0:(n-1)]^r$ such that for all $\mathbf{z}_{r+1} \neq \mathbf{t}_{r+1}$, it holds that $F_r(x, \mathbf{z}_{r+1}, s) = 0$. This can equivalently be expressed as

$$\bigcap_{\mathbf{z}_{r+1}} \mathrm{supp}\big(F_r(x, \mathbf{z}_{r+1}, s)\big) = \varnothing. \tag{13}$$

Next, since $f_{X_{r+1}}^{\mathbf{z}_{r+1}}(0) = 0$, for all $\mathbf{z}_{r+1}$, it follows from (13) that, for all $\mathbf{t}_{r+1} \in [0:(n-1)]^r$,

$$f_{X_{r+1}}^{\mathbf{t}_{r+1}}\big(F_r(x, \mathbf{t}_{r+1}, s)\big) > 0 \Rightarrow Z_{r+1}(x, s) := \sum_{\mathbf{z}_{r+1}} f_{X_{r+1}}^{\mathbf{z}_{r+1}}\big(F_r(x, \mathbf{z}_{r+1}, s)\big) = f_{X_{r+1}}^{\mathbf{t}_{r+1}}\big(F_r(x, \mathbf{t}_{r+1}, s)\big), \ x \in [0,1].$$

This concludes the proof. $\qquad\square$

Before proceeding, we need to introduce further notation. Specifically, let $P_1 = [a,b], P_2 = [c,d]$ be intervals in $[0,1]$, i.e., $0 \leq a < b \leq 1$ and $0 \leq c < d \leq 1$. Then, we define $P := P_1 \diamond P_2$ according to $P = [a + c(b-a), a + d(b-a)]$. Note that $|P| = |P_1||P_2|$. The $\diamond$-operation is associative in the sense of $(P_1 \diamond P_2) \diamond P_3 = P_1 \diamond (P_2 \diamond P_3)$. We further define the function $N(x, [a,b]) = a + x(b-a)$, $x \in [0,1]$, which rescales $[0,1]$ to the interval $[a,b]$, and note that

$$N(N(x, [a,b]), [c,d]) = N(x, [c,d] \diamond [a,b]). \tag{14}$$

We will also need the function $N^-(x, [a,b]) = b - x(b-a)$, $x \in [0,1]$, which, like $N(x, [a,b])$, rescales $[0,1]$ to the interval $[a,b]$, but does so in reverse manner, i.e., by mapping $x = 0$ to b and $x = 1$ to a. Additionally, we define the operator $S([a,b]) = [1-b, 1-a]$, for all $0 \leq a < b \leq 1$, which maps the interval $[a,b] \subseteq [0,1]$ to the interval $[1-b, 1-a]$. Note that S is cardinality-preserving, i.e., $|S([a,b])| = (b-a) = |[a,b]|$. Moreover, we have the relation

$$1 - N(x, S([a,b])) = N^-(x, [a,b]). \tag{15}$$

The next lemma establishes that the Z_r-functions indeed realize the per-bin histogram distributions constituting the desired target histogram distribution.

Lemma V.4. *Let* $\mathbf{z} = (z_1, z_2, \ldots, z_d) \in [0:(n-1)]^d$ *and* $\Delta_h = \left[\frac{h}{2^s}, \frac{h+1}{2^s}\right]$, $h \in [0:(2^s-1)]$. *Set* $c_{z_i} = \left[\frac{z_i}{n}, \frac{z_i+1}{n}\right]$, *for* $i \in [1:d]$. *Let* $\mathbf{z}_i = \mathbf{z}_{[1:(i-1)]}$ *and fix* $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]_n^d$ *with* $p_{X_1}^{\mathbf{z}_1} := p_{X_1}$ *and* $p_{X_i}^{\mathbf{z}_i} := p_{X_i|X_1,\ldots,X_{i-1}}\big(x_i|\mathbf{x}_{[1:(i-1)]} \in c_{\mathbf{z}_i}\big)$, *for* $i \in [2:d]$, *where* $p_{X_i}^{\mathbf{z}_i} \in \mathcal{E}[0,1]_n^1$, $i \in [1:d]$, *has weights* $w_k^{\mathbf{z}_i}$, *for* $k \in [0:(n-1)]$. *Let* $f_{X_i}^{\mathbf{z}_i}$ *be the piecewise linear function which, according to Corollary IV.4, satisfies* $f_{X_i}^{\mathbf{z}_i} \# U = p_{X_i}^{\mathbf{z}_i}$, $i \in [1:d]$. *Define* $P_r^{\mathbf{z}_i} = \left[\frac{1}{n}\sum_{k=0}^{r-1} w_k^{\mathbf{z}_i}, \frac{1}{n}\sum_{k=0}^{r} w_k^{\mathbf{z}_i}\right]$ *and* $P_r^{\mathbf{z}_i,h} = P_r^{\mathbf{z}_i} \diamond \Delta_h$, $r \in [1:(n-1)]$, *and* $P_0^{\mathbf{z}_i} = \left[0, \frac{w_0^{\mathbf{z}_i}}{n}\right]$, $P_0^{\mathbf{z}_i,h} = P_0^{\mathbf{z}_i} \diamond \Delta_h$. *Then, for every* $k \in [2:d]$, *for all* $\mathbf{z}_k \in [0:(n-1)]^{k-1}$ *and* $\mathbf{h}_k = (h_1, h_2, \ldots, h_{k-1}) \in [0:(2^s-1)]^{k-1}$, *it holds that*

$$Z_k(N(x, T_{k-1}), s) = f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_{k-1}), \mathbf{z}_k, s)) = f_{X_k}^{\mathbf{z}_k}(x), \ \textit{for} \ x \in [0,1] \ \textit{and} \ \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0,$$

and

$$Z_k(N(x, T_{k-1}), s) = f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_{k-1}), \mathbf{z}_k, s)) = f_{X_k}^{\mathbf{z}_k}(1-x), \ \textit{for} \ x \in [0,1] \ \textit{and} \ \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1,$$

where T_i, $i \in [1:d]$, *is defined recursively according to*

$$T_i = \begin{cases} T_{i-1} \diamond P_i^{\mathbf{z}}, & \textit{if } \sum_{\ell=1}^{i-1} h_\ell \in 2\mathbb{N}_0 \\ T_{i-1} \diamond S(P_i^{\mathbf{z}}), & \textit{if } \sum_{\ell=1}^{i-1} h_\ell \in 2\mathbb{N}_0 + 1 \end{cases},$$

for $i \in [2:d]$, *and initialized by* $T_1 = P_0^{\mathbf{z}} \diamond P_1^{\mathbf{z}}$, *with* $P_i^{\mathbf{z}} := P_{z_i}^{\mathbf{z}_i,h_i}$ *and* $P_0^{\mathbf{z}} = [0,1]$. *Moreover,* $|T_k| = \frac{1}{2^{sk}} p_{\mathbf{X}}(\mathbf{x}_{[1:k]} \in c_{\mathbf{z}_{k+1}})$, *for all* $k \in [1:(d-1)]$.

Proof. The proof is through induction across k. We start with the base case $k = 2$. Since the vectors $\mathbf{z}_2$ and $\mathbf{t}_2$ are, in fact, scalars, we can write $\mathbf{z}_2 = z_1$ and $\mathbf{t}_2 = t_1$. Fix $h_1 \in [0:(2^s - 1)]$, $z_1 \in [0:(n-1)]$, and note

13

that $T_1 = [0,1] \diamond P_1^{\mathbf{z}} = P_1^{\mathbf{z}} = P_{z_1}^{\mathbf{z}_1} \diamond \Delta_{h_1} = \left[\frac{1}{n} \sum_{k=0}^{z_1-1} w_k^{\mathbf{z}_1}, \frac{1}{n} \sum_{k=0}^{z_1} w_k^{\mathbf{z}_1} \right] \diamond \left[\frac{h_1}{2^s}, \frac{h_1+1}{2^s} \right] = \left[\frac{1}{n} \sum_{k=0}^{z_1-1} w_k^{\mathbf{z}_1} + \right.$

$\left. \frac{h_1 w_{z_1}^{\mathbf{z}_1}}{2^s n}, \frac{1}{n} \sum_{k=0}^{z_1-1} w_k^{\mathbf{z}_1} + \frac{(h_1+1) w_{z_1}^{\mathbf{z}_1}}{2^s n} \right]$. Further, note that $|T_1| = \frac{w_{z_1}^{\mathbf{z}_1}}{2^s n} = \frac{1}{2^s} p_{\mathbf{X}}(x_1 \in c_{z_1})$. By Corollary IV.4, $f_{X_1}^{\mathbf{z}_1}(x)$ is

linear on $P_{z_1}^{\mathbf{z}_1}$ with slope $\frac{1}{w_{z_1}^{\mathbf{z}_1}}$ and boundary points $f_{X_1}^{\mathbf{z}_1}(\frac{1}{n} \sum_{k=0}^{z_1-1} w_k^{\mathbf{z}_1}) = z_1/n$ and $f_{X_1}^{\mathbf{z}_1}(\frac{1}{n} \sum_{k=0}^{z_1} w_k^{\mathbf{z}_1}) = (z_1+1)/n$.
The explicit form of $f_{X_1}^{\mathbf{z}_1}(x)$ on $P_{z_1}^{\mathbf{z}_1}$ follows from the remark after Corollary IV.4 as

$$f_{X_1}^{\mathbf{z}_1}(x) = \frac{x}{w_{z_1}^{\mathbf{z}_1}} - \frac{\sum_{i=0}^{z_1-1} w_i^{\mathbf{z}_1}}{n w_{z_1}^{\mathbf{z}_1}} + \frac{z_1}{n}, \quad x \in P_{z_1}^{\mathbf{z}_1}.$$

Next, since $N(x, T_1) = \frac{1}{n} \sum_{k=0}^{z_1-1} w_k^{\mathbf{z}_1} + \frac{(x+h_1) w_{z_1}^{\mathbf{z}_1}}{2^s n}$, noting that $T_1 \subset P_{z_1}^{\mathbf{z}_1}$, we obtain

$$f_{X_1}^{\mathbf{z}_1}(N(x,T_1)) = \frac{x+h_1}{2^s n} + \frac{z_1}{n}, \quad x \in [0,1], \tag{16}$$

and, hence, for $x \in [0,1]$,

$$
\begin{aligned}
f_{X_2}^{(z_1)}\big(F_1(N(x,T_1), z_1, s)\big) &= f_{X_2}^{(z_1)}\big(g_s\big(n f_{X_1}^{\mathbf{z}_1}(N(x,T_1)) - z_1\big)\big) \\
&= f_{X_2}^{(z_1)}(g_s((x+h_1)2^{-s})) \\
&\overset{(a)}{=} \sum_{j=0}^{2^{s-1}-1} f_{X_2}^{(z_1)}(g(2^{s-1}(x+h_1)2^{-s} - j)) \\
&= \sum_{j=0}^{2^{s-1}-1} f_{X_2}^{(z_1)}(g(x/2 + h_1/2 - j)) \\
&\overset{(b)}{=} f_{X_2}^{(z_1)}(g(x/2 + h_1/2 - \lfloor h_1/2 \rfloor)) \\
&\overset{(c)}{=} \begin{cases} f_{X_2}^{(z_1)}(x), & \text{if } h_1 \in 2\mathbb{N}_0, \\ f_{X_2}^{(z_1)}(1-x), & \text{if } h_1 \in 2\mathbb{N}_0 + 1, \end{cases}
\end{aligned}
\tag{17}
$$

where we used Lemma III.1 in (a), the fact that $g(x/2 + h_1/2 - j) = 0$, for all $x \in [0,1]$, for $j \neq \lfloor h_1/2 \rfloor$ in (b), and $h_1/2 - \lfloor h_1/2 \rfloor = 0$ for $h_1 \in 2\mathbb{N}_0$ and $h_1/2 - \lfloor h_1/2 \rfloor = 1/2$ for $h_1 \in 2\mathbb{N}_0 + 1$ along with $g(x) = g(1-x)$, for $x \in [0,1]$, in (c). Finally, as by Corollary IV.4, $f_{X_2}^{(z_1)}(x) > 0$, for all $x \in (0,1]$, it follows from (17) that $f_{X_2}^{(z_1)}\big(F_1(N(x,T_1), z_1, s)\big) > 0$, for all $x \in (0,1]$ for $h_1 \in 2\mathbb{N}_0$, and for all $x \in [0,1)$ for $h_1 \in 2\mathbb{N}_0 + 1$. Application of Lemma V.3 then yields

$$Z_2(N(x,T_1), s) = f_{X_2}^{(z_1)}\big(F_1(N(x,T_1), z_1, s)\big), \tag{18}$$

for all $x \in (0,1]$ for $h_1 \in 2\mathbb{N}_0$, and for all $x \in [0,1)$ for $h_1 \in 2\mathbb{N}_0 + 1$. To see that (18) holds for $x = 0$ and $h_1 \in 2\mathbb{N}_0$, simply note that $Z_2(N(0,T_1), s) = \sum_{z_1} f_{X_2}^{(z_1)}(F_1(N(0,T_1), z_1, s))$ and $F_1(N(0,T_1), z_1, s) = g_s(h_1/2^s) = 0$, which thanks to $f_{X_2}^{(z_1)}(0) = 0$ implies $Z_2(N(0,T_1), s) = f_{X_2}^{(z_1)}(F_1(N(0,T_1), z_1, s)) = 0$. The case $x = 1$ and $h_1 \in 2\mathbb{N}_0 + 1$ follows along the exact same lines noting that $F_1(N(1,T_1), z_1, s) = g_s((h_1+1)/2^s) = 0$. This finalizes the proof of the base case.

The proof of the induction step largely follows the arguments underlying the proof of the base case. Fix $k \in \mathbb{N}$, with $k \geq 2$, and assume that for all $\mathbf{z}_k = (z_1, z_2, \ldots, z_{k-1}) \in [0{:}(n-1)]^{k-1}$ and $\mathbf{h}_k = (h_1, h_2, \ldots, h_{k-1}) \in [0{:}(2^s - 1)]^{k-1}$, it holds that

$$Z_k(N(x, T_{k-1}), s) = f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_{k-1}), \mathbf{z}_k, s)) = f_{X_k}^{\mathbf{z}_k}(x), \quad \text{for } x \in [0,1] \text{ and } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0,$$

and

$$Z_k(N(x, T_{k-1}), s) = f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_{k-1}), \mathbf{z}_k, s)) = f_{X_k}^{\mathbf{z}_k}(1-x), \quad \text{for } x \in [0,1] \text{ and } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1,$$

14

with $|T_{k-1}| = \frac{1}{2^{s(k-1)}} p_{\mathbf{X}}(\mathbf{x}_{[1:(k-1)]} \in c_{\mathbf{z}_k})$. Fix $\mathbf{h}_{k+1} = (h_1, h_2, \ldots, h_k) \in [0:(2^s-1)]^k$ and $\mathbf{z}_{k+1} = (z_1, z_2, \ldots, z_k) \in [0:(n-1)]^k$. Consider $Z_k(N(x, T_{k-1}), s) = f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_{k-1}), \mathbf{z}_k, s))$ on the interval

$$P_k^{\mathbf{z}} = P_{z_k}^{\mathbf{z}_k} \diamond \Delta_{h_k} = \left[\frac{1}{n} \sum_{j=0}^{z_k-1} w_j^{\mathbf{z}_k} + \frac{h_k w_{z_k}^{\mathbf{z}_k}}{2^s n}, \frac{1}{n} \sum_{j=0}^{z_k-1} w_j^{\mathbf{z}_k} + \frac{(h_k+1) w_{z_k}^{\mathbf{z}_k}}{2^s n} \right].$$

We first note that

$$T_k = \begin{cases} T_{k-1} \diamond P_k^{\mathbf{z}}, & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 \\ T_{k-1} \diamond S(P_k^{\mathbf{z}}), & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases},$$

and

$$\begin{aligned} |T_k| &= \begin{cases} |T_{k-1}||P_k^{\mathbf{z}}|, & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 \\ |T_{k-1}||S(P_k^{\mathbf{z}})|, & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases} \\ &= |T_{k-1}| \frac{w_{z_k}^{\mathbf{z}_k}}{2^s n} \\ &= \frac{1}{2^{s(k-1)}} p_{\mathbf{X}}(\mathbf{x}_{[1:(k-1)]} \in c_{\mathbf{z}_k}) \frac{1}{2^s} p_{X_k | X_1, \ldots, X_{k-1}}(x_k \in c_{z_k} | \mathbf{x}_{[1:(k-1)]} \in c_{\mathbf{z}_k}) \\ &= \frac{1}{2^{sk}} p_{\mathbf{X}}(\mathbf{x}_{[1:k]} \in c_{\mathbf{z}_{k+1}}). \end{aligned}$$

We first provide the proof for the case $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0$. By Corollary IV.4, $f_{X_k}^{\mathbf{z}_k}(x)$ is linear on $P_{z_k}^{\mathbf{z}_k}$ with slope $1/w_{z_k}^{\mathbf{z}_k}$ and boundary points $f_{X_k}^{\mathbf{z}_k}(\frac{1}{n} \sum_{j=0}^{z_k} w_j^{\mathbf{z}_k}) = z_k/n$ and $f_{X_k}^{\mathbf{z}_k}(\frac{1}{n} \sum_{j=0}^{z_k} w_j^{\mathbf{z}_k}) = (z_k + 1)/n$. The explicit form of $f_{X_k}^{\mathbf{z}_k}$ on $P_{z_k}^{\mathbf{z}_k}$ follows from the remark after Corollary IV.4 as

$$f_{X_k}^{\mathbf{z}_k}(x) = \frac{x}{w_{z_k}^{\mathbf{z}_k}} - \frac{\sum_{j=0}^{z_k-1} w_j^{\mathbf{z}_k}}{n w_{z_k}^{\mathbf{z}_k}} + \frac{z_k}{n}.$$

Using $N(x, T_k) = N(N(x, P_k^{\mathbf{z}}), T_{k-1})$, which is thanks to (14), in the induction assumption (for $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0$), we get

$$\begin{aligned} f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_k), \mathbf{z}_k, s)) &= f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(N(x, P_k^{\mathbf{z}}), T_{k-1}), \mathbf{z}_k, s)) \\ &= f_{X_k}^{\mathbf{z}_k}(N(x, P_k^{\mathbf{z}})) = \frac{x + h_k}{2^s n} + \frac{z_k}{n}, \quad x \in [0, 1]. \end{aligned} \tag{19}$$

Next, for $x \in [0, 1]$, it follows from (9) and (19) that

$$\begin{aligned} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(F_k(N(x, T_k), \mathbf{z}_{k+1}, s)) &= f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g_s(n f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_k), \mathbf{z}_k, s)) - z_k)) \\ &= f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g_s((x + h_k) 2^{-s})) \\ &\overset{(a)}{=} \sum_{j=0}^{2^{s-1}-1} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(2^{s-1}(x + h_k) 2^{-s} - j)) \\ &= \sum_{j=0}^{2^{s-1}-1} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(x/2 + h_k/2 - j)) \\ &\overset{(b)}{=} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(x/2 + h_k/2 - \lfloor h_k/2 \rfloor)) \\ &\overset{(c)}{=} \begin{cases} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x), & \text{if } h_k \in 2\mathbb{N}_0, \\ f_{X_{k+1}}^{\mathbf{z}_{k+1}}(1 - x), & \text{if } h_k \in 2\mathbb{N}_0 + 1, \end{cases} \\ &\overset{(d)}{=} \begin{cases} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x), & \text{if } \sum_{i=1}^{k} h_i \in 2\mathbb{N}_0, \\ f_{X_{k+1}}^{\mathbf{z}_{k+1}}(1 - x), & \text{if } \sum_{i=1}^{k} h_i \in 2\mathbb{N}_0 + 1, \end{cases} \end{aligned} \tag{20}$$

where we used Lemma III.1 in (a), the fact that $g(x/2 + h_k/2 - j) = 0$, for all $x \in [0, 1]$, for $j \neq \lfloor h_k/2 \rfloor$ in (b), $h_k/2 - \lfloor h_k/2 \rfloor = 0$ for $h_k \in 2\mathbb{N}_0$ and $h_k/2 - \lfloor h_k/2 \rfloor = 1/2$ for $h_k \in 2\mathbb{N}_0 + 1$ along with $g(x) = g(1 - x)$,

for $x \in [0, 1]$, in (c), and $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0$ in (d). Finally, as by Corollary IV.4, $f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x) > 0$, for all $x \in (0, 1]$, it follows from (20) that $f_{X_{k+1}}^{\mathbf{z}_{k+1}}\big(F_k(N(x, T_k), \mathbf{z}_{k+1}, s)\big) > 0$, for all $x \in (0, 1]$ for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0$, and for all $x \in [0, 1)$ for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0 + 1$. Application of Lemma V.3 then yields

$$Z_{k+1}(N(x, T_k), s) = f_{X_{k+1}}^{\mathbf{z}_{k+1}}\big(F_k(N(x, T_k), \mathbf{z}_{k+1}, s)\big),$$

for all $x \in (0, 1]$ for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0$, and for all $x \in [0, 1)$ for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0 + 1$. The boundary cases i) $x = 0$ and $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0$ and ii) $x = 1$ and $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0 + 1$ follow along the same lines as in the base case upon noting that $F_k(N(0, T_k), \mathbf{z}_{k+1}, s) = g_s(h_k/2^s)$ and $F_k(N(1, T_k), \mathbf{z}_{k+1}, s) = g_s((h_k + 1)/2^s)$.

We proceed to the proof for the case $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1$. Using (15) and $N(x, T_k) = N(N(x, S(P_k^{\mathbf{z}})), T_{k-1})$, which is thanks to (14), in the induction assumption (for $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1$), we get

$$
\begin{aligned}
f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_k), \mathbf{z}_k, s)) &= f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(N(x, S(P_k^{\mathbf{z}})), T_{k-1}), \mathbf{z}_k, s)) \\
&= f_{X_k}^{\mathbf{z}_k}(1 - N(x, S(P_k^{\mathbf{z}}))) \\
&= f_{X_k}^{\mathbf{z}_k}(N^-(x, P_k^{\mathbf{z}})) \\
&= \frac{h_k + 1 - x}{2^s n} + \frac{z_k}{n}, \quad x \in [0, 1].
\end{aligned}
\tag{21}
$$

Next, for $x \in [0, 1]$, it follows from (9) and (21) that

$$
\begin{aligned}
f_{X_{k+1}}^{\mathbf{z}_{k+1}}\big(F_k(N(x, T_k), \mathbf{z}_{k+1}, s)\big) &= f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g_s(n f_{X_k}^{\mathbf{z}_k}(F_{k-1}(N(x, T_k), \mathbf{z}_k, s)) - z_k)) \\
&= f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g_s((h_k + 1 - x)2^{-s})) \\
&\overset{(a)}{=} \sum_{j=0}^{2^{s-1}-1} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(2^{s-1}(h_k + 1 - x)2^{-s} - j)) \\
&= \sum_{j=0}^{2^{s-1}-1} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(h_k/2 + 1/2 - x/2 - j)) \\
&\overset{(b)}{=} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(g(h_k/2 + 1/2 - x/2 - \lfloor h_k/2 \rfloor)) \\
&\overset{(c)}{=} \begin{cases} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(1 - x), & \text{if } h_k \in 2\mathbb{N}_0, \\ f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x), & \text{if } h_k \in 2\mathbb{N}_0 + 1, \end{cases} \\
&\overset{(d)}{=} \begin{cases} f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x), & \text{if } \sum_{i=1}^{k} h_i \in 2\mathbb{N}_0, \\ f_{X_{k+1}}^{\mathbf{z}_{k+1}}(1 - x), & \text{if } \sum_{i=1}^{k} h_i \in 2\mathbb{N}_0 + 1, \end{cases}
\end{aligned}
\tag{22}
$$

where we used Lemma III.1 in (a), the fact that $g(h_k/2 + 1/2 - x/2 - j) = 0$, for all $x \in [0, 1]$, for $j \neq \lfloor h_k/2 \rfloor$ in (b), $h_k/2 - \lfloor h_k/2 \rfloor = 0$ for $h_k \in 2\mathbb{N}_0$ and $h_k/2 - \lfloor h_k/2 \rfloor = 1/2$ for $h_k \in 2\mathbb{N}_0 + 1$ along with $g(x) = g(1 - x)$, for $x \in [0, 1]$, in (c), and $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1$ in (d).

Finally, as by Corollary IV.4, $f_{X_{k+1}}^{\mathbf{z}_{k+1}}(x) > 0$, for all $x \in (0, 1]$, it follows from (22) that $f_{X_{k+1}}^{\mathbf{z}_{k+1}}\big(F_k(N(x, T_k), \mathbf{z}_{k+1}, s)\big) > 0$, for all $x \in (0, 1]$ for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0$, and for all $x \in [0, 1)$ for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0 + 1$. Application of Lemma V.3 then yields

$$Z_{k+1}(N(x, T_k), s) = f_{X_{k+1}}^{\mathbf{z}_{k+1}}\big(F_k(N(x, T_k), \mathbf{z}_{k+1}, s)\big),$$

for all $x \in (0, 1]$ for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0$, and for all $x \in [0, 1)$ for $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0 + 1$. The boundary cases i) $x = 0$ and $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0$ and ii) $x = 1$ and $\sum_{i=1}^{k} h_i \in 2\mathbb{N}_0 + 1$ follow along the same lines as in the base case upon noting that $F_k(N(0, T_k), \mathbf{z}_{k+1}, s) = g_s((h_k + 1)/2^s)$ and $F_k(N(1, T_k), \mathbf{z}_{k+1}, s) = g_s(h_k/2^s)$.

This concludes the proof of the induction step and thereby the overall proof. $\qquad\square$

We continue with a corollary to Lemma V.4 complementing the results on $|T_k|, k \in [1 : (d - 1)]$, by the corresponding expression for $|T_d|$ and specifying the range of the Z_r-functions on the domain T_d.

Corollary V.5. *Let* $\mathbf{z} = (z_1, z_2, \ldots, z_d) \in [0:(n-1)]^d$ *and* $\mathbf{z}_i = \mathbf{z}_{[1:(i-1)]}$. *Fix* $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]_n^d$, *and for all* $\mathbf{h}_d = (h_1, h_2, \ldots, h_d) \in [0:(2^s-1)]^d$, *let* $T_k, k \in [1:d]$, *be defined as in Lemma V.4. Then, it holds that* $|T_d| = \frac{1}{2^{sd}} p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}})$. *Moreover, for every* $k \in [1:d]$, *for all* $x \in T_d$, $Z_k(x,s) \in \left[\frac{z_k}{n} + \frac{h_k}{2^s n}, \frac{z_k}{n} + \frac{h_k+1}{2^s n}\right]$.

Proof. We first prove the statement on $|T_d|$ and start by noting that, owing to Lemma V.4,

$$|T_{d-1}| = \frac{1}{2^{s(d-1)}} p_{\mathbf{X}}(\mathbf{x}_{[1:(d-1)]} \in c_{\mathbf{z}_d}).$$

With

$$P_{z_d}^{\mathbf{z}_d, h_d} = P_{z_d}^{\mathbf{z}_d} \diamond \Delta_{h_d} = \left[\frac{1}{n}\sum_{j=0}^{z_d-1} w_j^{\mathbf{z}_d} + \frac{h_d w_{z_d}^{\mathbf{z}_d}}{2^s n}, \frac{1}{n}\sum_{j=0}^{z_d-1} w_j^{\mathbf{z}_d} + \frac{(h_d+1)w_{z_d}^{\mathbf{z}_d}}{2^s n}\right]$$

and $|P_{z_d}^{\mathbf{z}_d, h_d}| = |S(P_{z_d}^{\mathbf{z}_d, h_d})|$, we get

$$
\begin{aligned}
|T_d| &= \begin{cases} |T_{d-1}||P_{z_d}^{\mathbf{z}_d, h_d}|, & \text{if } \sum_{i=1}^{d-1} h_i \in 2\mathbb{N}_0 \\ |T_{d-1}||S(P_{z_d}^{\mathbf{z}_d, h_d})|, & \text{if } \sum_{i=1}^{d-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases} \\
&= \frac{1}{2^{s(d-1)}} p_{\mathbf{X}}(\mathbf{x}_{[1:(d-1)]} \in c_{\mathbf{z}_d}) \frac{w_{z_d}^{\mathbf{z}_d}}{2^s n} \\
&= \frac{1}{2^{s(d-1)}} p_{\mathbf{X}}(\mathbf{x}_{[1:(d-1)]} \in c_{\mathbf{z}_d}) \frac{1}{2^s} p_{X_d | X_1, \ldots, X_{d-1}}(x_d \in c_{z_d} | \mathbf{x}_{[1:(d-1)]} \in c_{\mathbf{z}_d}) \\
&= \frac{1}{2^{sd}} p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}}).
\end{aligned}
$$

This establishes the first statement.

To prove the second statement, we first note that, for $k = 1$, by (16), $Z_1(x,s) \in \left[\frac{z_1}{n} + \frac{h_1}{2^s n}, \frac{z_1}{n} + \frac{h_1+1}{2^s n}\right]$, for all $x \in T_1$. Next, for every $k \in [2:d]$, for all T_{k-1}, by Lemma V.4, it holds that

$$Z_k(N(x, T_{k-1}), s) = \begin{cases} f_{X_k}^{\mathbf{z}_k}(x), & \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 \\ f_{X_k}^{\mathbf{z}_k}(1-x), & \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases},$$

for all $x \in [0,1]$. Now, arbitrarily fix $z_k \in [0:(n-1)], h_k \in [0:(2^s-1)]$ and consider

$$T_k = \begin{cases} T_{k-1} \diamond P_k^{\mathbf{z}}, & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 \\ T_{k-1} \diamond S(P_k^{\mathbf{z}}), & \text{if } \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases}$$

with $P_k^{\mathbf{z}} = P_{z_k}^{\mathbf{z}_k, h_k}$. With (14), this yields, for all $x \in [0,1]$,

$$Z_k(N(x, T_k), s) = \begin{cases} Z_k(N(N(x, P_k^{\mathbf{z}}), T_{k-1}), s) = f_{X_k}^{\mathbf{z}_k}(N(x, P_k^{\mathbf{z}})), & \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 \\ Z_k(N(N(x, S(P_k^{\mathbf{z}})), T_{k-1}), s) = f_{X_k}^{\mathbf{z}_k}(1 - N(x, S(P_k^{\mathbf{z}}))), & \sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1 \end{cases}.$$

Now, by (19), it follows for $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0$ that

$$f_{X_k}^{\mathbf{z}_k}(N(x, P_k^{\mathbf{z}})) = \frac{x + h_k}{2^s n} + \frac{z_k}{n} \in \left[\frac{z_k}{n} + \frac{h_k}{2^s n}, \frac{z_k}{n} + \frac{h_k+1}{2^s n}\right], \text{ for } x \in [0,1],$$

and analogously, for $\sum_{i=1}^{k-1} h_i \in 2\mathbb{N}_0 + 1$, by (21),

$$f_{X_k}^{\mathbf{z}_k}(1 - N(x, S(P_k^{\mathbf{z}}))) = f_{X_k}^{\mathbf{z}_k}(N^-(x, P_k^{\mathbf{z}})) = \frac{h_k + 1 - x}{2^s n} + \frac{z_k}{n} \in \left[\frac{z_k}{n} + \frac{h_k}{2^s n}, \frac{z_k}{n} + \frac{h_k+1}{2^s n}\right], \text{ for } x \in [0,1].$$

We have hence shown that, for all $k \in [1:d]$, $Z_k(x,s) \in \left[\frac{z_k}{n} + \frac{h_k}{2^s n}, \frac{z_k}{n} + \frac{h_k+1}{2^s n}\right]$, for all $x \in T_k$. The proof is completed upon noting that $T_d \subseteq T_k$, for all $k \in [1:d]$. $\square$

We are now ready to state the main result of this section, namely that the piecewise linear map

$$M : x \to (Z_1(x,s), Z_2(x,s), \ldots, Z_d(x,s))$$

17

transports a 1-dimensional uniform distribution in a space-filling manner to an arbitrarily close approximation of any high-dimensional histogram distribution.

Theorem V.6. *For every distribution $p_{\mathbf{X}}(\mathbf{x}) \in \mathcal{E}[0,1]_n^d$, the corresponding transport map*

$$M : x \to (Z_1(x,s), Z_2(x,s), \ldots, Z_d(x,s)) \tag{23}$$

satisfies

$$W(M\#U, p_{\mathbf{X}}) \leq \frac{\sqrt{d}}{n2^s}.$$

Proof. Let $\mathbf{z} = (z_1, z_2, \ldots, z_d) \in [0:(n-1)]^d$, $\Delta_h = \left[\frac{h}{2^s}, \frac{h+1}{2^s}\right]$ with $h \in [0:(2^s-1)]$, and $\mathbf{h} = (h_1, h_2, \ldots, h_d) \in [0:(2^s-1)]^d$. With $c_{z_i} = [\frac{z_i}{n}, \frac{z_i+1}{n}]$, $i \in [1:d]$, let $c_{\mathbf{z}}^{\mathbf{h}} = \bigtimes_{i=1}^d (c_{z_i} \diamond \Delta_{h_i})$. Let T_d be defined as in Lemma V.4. By Corollary V.5, $M : T_d \to c_{\mathbf{z}}^{\mathbf{h}}$ and $|T_d| = \frac{1}{2^{sd}} p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}})$. We hence get $(M\#U)(\mathbf{x} \in c_{\mathbf{z}}^{\mathbf{h}}) = |T_d| = \frac{1}{2^{sd}} p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}})$. This establishes that the map M transports probability mass $\frac{1}{2^{sd}} p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}})$ to the cube $c_{\mathbf{z}}^{\mathbf{h}}$ of volume $\left(\frac{1}{n2^s}\right)^d$, for all $\mathbf{z}$. As $p_{\mathbf{X}}$ is a histogram distribution, it is uniformly distributed on its constituent cubes $c_{\mathbf{z}}$, which, in turn, implies that the amount of probability mass it exhibits on each subcube $c_{\mathbf{z}}^{\mathbf{h}}$ of $c_{\mathbf{z}}$ is given by $\frac{1}{2^{sd}} p_{\mathbf{X}}(\mathbf{x} \in c_{\mathbf{z}})$. The map M, when pushing forward U, therefore transports exactly the right amount of probability mass to each cube $c_{\mathbf{z}}^{\mathbf{h}}$ for a coupling between $p_{\mathbf{X}}$ and $M\#U$ to exist. Combining this with $\|\mathbf{x} - \mathbf{y}\| \leq \frac{\sqrt{d}}{n2^s}$, for all points $\mathbf{x}, \mathbf{y}$ in a d-dimensional cube of side length $n^{-1}2^{-s}$, it follows from Definition II.2 that

$$W(M\#U, p_{\mathbf{X}}) \leq \frac{\sqrt{d}}{n2^s}. \qquad \square$$

Theorem V.6 was proven in [19] for $d = 2$. We remark that a space-filling approach for increasing distribution dimensionality was first described by Bailey and Telgarsky in [2]. Specifically, the construction in [2] generates uniform target distributions of arbitrary dimension based on the transport map $M : x \to (x, g_s(x), g_{2s}(x), \ldots)$. The generalization introduced in this paper is capable of producing arbitrary histogram target distributions through space-filling transport maps that build on several key ideas, the first two of which are best illustrated by revisiting the 2-dimensional case with corresponding transport map $M : x \to (f_{X_1}^{\mathbf{z}_1}(x), \sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(n f_{X_1}^{\mathbf{z}_1}(x) - i)))$. First, M in its second component composes the function $\sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(nx - i))$ with its first component $f_{X_1}^{\mathbf{z}_1}(x)$. Formally, this idea is also present in the Bailey-Telgarsky map, where the second component $g_s(x)$ can be interpreted as a trivial composition of $g_s(\cdot)$ with the first component, x. It is, in fact, this composition idea that leads to the space-filling property. Second, $\sum_{i=0}^{n-1} f_{X_2}^{(i)}(g_s(n f_{X_1}^{\mathbf{z}_1}(x) - i))$ yields localization through squeezing and shifting of the $f_{X_2}^{(i)}$. This idea allows to realize different marginal distributions for different horizontal histogram bins (see the rightmost subplot in Fig. 6) and is not present in the Bailey-Telgarsky construction as, owing to the target distribution being uniform, there is no concept of histogram distributions. Taken together the two ideas just described allow to generate arbitrary marginal histogram distributions $p_{X_2|X_1}(x_2|x_1)$, which are then combined—through the chain rule—with the histogram distribution $p_{X_1}(x_1)$ to the overall target histogram distribution $p_{X_1,X_2}(x_1, x_2)$.

A further idea underlying our transport map construction becomes transparent in the general d-dimensional case. Specifically, in taking the space-filling idea to higher dimensions, we note that in the Bailey-Telgarsky map $M : x \to (x, g_s(x), g_{2s}(x), \ldots)$, the third component, $g_{2s}(x)$ can actually be interpreted as a composition of $g_s(\cdot)$ with the second component $g_s(x)$, simply as $g_s(g_s(x)) = g_{2s}(x)$. Likewise, as already noted in the previous paragraph, the second component, $g_s(x)$, is a composition of $g_s(\cdot)$ with the first component, x. This insight informs the recursive definition of the F_r-functions according to (9), which, modulo the shaping by the localized $f_{X_r}^{\mathbf{z}_r}$-functions, can be seen to exhibit this g_s-composition property as well. The $Z_r(x,s)$-functions constituting the components of our transport map (23) are then obtained by applying the localization idea as described above for the 2-dimensional case. There is, however, an important difference between localization in the 2-dimensional case and in the general d-dimensional case. This is best seen by inspecting the 3-dimensional case illustrated in Figure 7. Specifically, whereas in the 2-dimensional case the F_r-functions are contiguously supported (see subplot (f)), in the 3-dimensional case, as illustrated in subplot (d), the support sets are disjointed, but exhibit a periodic pattern. Going to higher dimensions yields a fractal-like support set picture. We emphasize that this support set structure is a consequence of interlacing the self-compositions of the g_s-functions with the localized per-bin histogram-distribution shaping functions $f_{X_r}^{\mathbf{z}_r}$.

We finally note that the transport map M in Theorem V.6 can be interpreted as a transport operator in the sense of optimal transport theory [20], [25], with the source distribution being 1-dimensional and the target-distribution d-dimensional. What is special here is that the transport operator acts between spaces of different dimensions and does so in a space-filling manner [18].

This section is concerned with the realization of the transport map M by ReLU networks. In particular, we shall consider networks with quantized weights, for three reasons. First, in practice network weights can not be stored as real numbers on a computer, but rather have to be encoded with a finite number of bits. Second, we want to convince ourselves that the space-filling property of the transport map, brittle as it seems, is, in fact, not dependent on the network weights being real numbers. Third, we will be able to develop a relationship, presented in Section IX, between the complexity of target distributions and the complexity of the ReLU networks realizing the corresponding transport maps. Specifically, complexity will be quantified through the number of bits needed to encode the distribution and the network, respectively, to within a prescribed accuracy.

We will see that ReLU networks with quantized weights generate histogram distributions with quantized weights, referred to as quantized histogram distributions in the following. In Section VII, we will then study the approximation of general distributions by quantized histogram distributions. Finally, in Section VIII, we put everything together and characterize the error incurred when approximating arbitrary target distributions by the transportation of a 1-dimensional uniform distribution through a ReLU network with quantized weights.

Before proceeding, we need to define quantized histogram distributions and quantized networks. We start with scalar distributions.

Definition VI.1. *Let $\delta = 1/A$, for some $A \in \mathbb{N}$. A random variable X is said to have a δ-quantized histogram distribution of resolution n on $[0, 1]$, denoted as $X \sim \tilde{\mathcal{E}}_\delta[0, 1]_n^1$, if its pdf is given by*

$$p(x) = \sum_{k=0}^{n-1} w_k \chi_{[k/n, (k+1)/n]}(x), \quad \sum_{k=0}^{n-1} w_k = n,$$
$$w_k = \delta m_k > 0, \quad m_k \in \mathbb{N}, \quad \text{for all} \quad k \in [0 : (n-1)].$$

We extend this definition to random vectors by saying that a random vector has a δ-quantized histogram distribution, if all its conditional (1-dimensional) distributions $p_{X_i}^{\mathbf{z}_i}$ are δ-quantized histogram distributions.

Definition VI.2. *Let $\delta = 1/A$, for some $A \in \mathbb{N}$. A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)^\top$ is said to have a δ-quantized histogram distribution of resolution n on the d-dimensional unit cube, denoted as $\mathbf{X} \sim \tilde{\mathcal{E}}_\delta[0, 1]_n^d$, if $\mathbf{X} \sim \mathcal{E}[0, 1]_n^d$ with $p_{X_i}^{\mathbf{z}_i} \in \tilde{\mathcal{E}}_\delta[0, 1]_n^1$, for every $i \in [1 : d]$, for all $\mathbf{z}_i$.*

We continue with the definition of quantized ReLU networks.

Definition VI.3. *For $\delta > 0$, we say that a ReLU network is δ-quantized if each of its weights is of one of the following two types. A weight w is of Type 1 if $w \in (\delta\mathbb{Z} \cap [-1/\delta, 1/\delta])$ and of Type 2 if $\frac{1}{w} \in (\delta\mathbb{Z} \cap [-1/\delta, 1/\delta])$.*

Formally, the goal of this section is to find, for fixed $p_\mathbf{X} \in \tilde{\mathcal{E}}_\delta[0, 1]_n^d$, a quantized ReLU network Φ such that $\Phi \# U$ approximates $p_\mathbf{X}$ to within a prescribed accuracy. To this end, we start with an auxiliary lemma, which constructs the building blocks of such networks.

Lemma VI.4. *For every δ-quantized $p_\mathbf{X} \in \tilde{\mathcal{E}}_\delta[0, 1]_n^d$ with $d > 1$, the map $M^r : \mathbb{R}^{n^r + r} \to \mathbb{R}^{n^{r+1} + r + 1}$, $r \in [0 : (d-1)]$, defined as*

$$M^0 : F_0(x, \mathbf{z}_1, s) \to \Big(F_1(x, \mathbf{z}_2^1, s), F_1(x, \mathbf{z}_2^2, s), \ldots, F_1(x, \mathbf{z}_2^n, s), Z_1(x, s) \Big),$$

and, for $r \in [1 : (d-1)]$,

$$M^r : \Big(F_r(x, \mathbf{z}_{r+1}^1, s), F_r(x, \mathbf{z}_{r+1}^2, s), \ldots, F_r(x, \mathbf{z}_{r+1}^{n^r}, s), Z_1(x, s), Z_2(x, s), \ldots, Z_r(x, s) \Big)$$
$$\to \Big(F_{r+1}(x, \mathbf{z}_{r+2}^1, s), F_{r+1}(x, \mathbf{z}_{r+2}^2, s), \ldots, F_{r+1}(x, \mathbf{z}_{r+2}^{n^{r+1}}, s), Z_1(x, s), Z_2(x, s), \ldots, Z_{r+1}(x, s) \Big),$$

is realizable through a Δ-quantized ReLU network $\Psi^{M^r} \in \mathcal{N}_{n^r+r, n^{r+1}+r+1}$ with $\mathcal{M}(\Psi^{M^r}) = \mathcal{O}(n^{r+2} + sn^{r+1})$ and $\mathcal{L}(\Psi^{M^r}) = s + 3$. Here, $\Delta = \frac{\delta}{n}$ and the vectors $\mathbf{z}_r^i \in [0:(n-1)]^{r-1}$, $i \in [1:n^{r-1}]$, are in natural order[3] with respect to i.

Proof. We start with auxiliary results needed in the proof and then proceed to establish the statement for the cases $r = 0$ and $r \geq 1$ separately. According to Corollary IV.4, for every $k \in [1:d]$, for all $\mathbf{z}_k \in [0:(n-1)]^{k-1}$, $f_{X_k}^{\mathbf{z}_k}(x)$ can be realized through a ReLU network $\Phi^{\mathbf{z}_k} : \mathbb{R} \to \mathbb{R} \in \mathcal{N}_{1,1}$ given by

$$\Phi^{\mathbf{z}_k} : x \to \frac{1}{w_0}\rho(x) + \sum_{i=1}^{n-1} \left(\frac{1}{w_i} - \frac{1}{w_{i-1}}\right)\rho\left(x - \frac{1}{n}\sum_{j=0}^{i-1} w_j\right),$$

and satisfying $\mathcal{M}(\Phi^{\mathbf{z}_k}) \leq 4n - 2$, $\mathcal{L}(\Phi^{\mathbf{z}_k}) = 2$. For $\Delta = \frac{\delta}{n}$, the network $\Phi^{\mathbf{z}_k}$ is Δ-quantized with the weights $\frac{1}{w_0}$, $\frac{1}{w_i}$, and $\frac{1}{w_{i-1}}$ of Type 2, and the weights $\frac{1}{n}\sum_{j=0}^{i-1} w_j$ of Type 1. The networks $\Phi_i^{\mathbf{z}_k}(x)$ implementing $\left(nf_{X_k}^{\mathbf{z}_k}(x) - i\right)$ are in $\mathcal{N}_{1,1}$ and have $\mathcal{M}(\Phi_i^{\mathbf{z}_k}) \leq 4n - 1$, $\mathcal{L}(\Phi_i^{\mathbf{z}_k}) = 2$, with their weights all of either Type 1 or Type 2 w.r.t. Δ-quantization. The network $\Psi_g^s(x)$ realizing $g_s(x)$ (see Section III) is in $\mathcal{N}_{1,1}$ with $\mathcal{M}(\Psi_g^s) = 11s - 3$, $\mathcal{L}(\Psi_g^s) = s + 1$, and with all its weights in $\{-4, -2, -1, 1, 2, 4\}$, which are, again, of Type 1 w.r.t. Δ-quantization. It follows from [9, Lemma II.3] that the networks $\Psi_{i,s}^{\mathbf{z}_k} = \Psi_g^s(\Phi_i^{\mathbf{z}_k})$ are in $\mathcal{N}_{1,1}$ with $\mathcal{M}(\Psi_{i,s}^{\mathbf{z}_k}) \leq 8n + 22s - 8$ and $\mathcal{L}(\Psi_{i,s}^{\mathbf{z}_k}) = s + 3$.

We are now ready to prove the statement for $r = 0$. Here, $M^0 : \mathbb{R} \to \mathbb{R}^{n+1}$ with

$$M^0 : F_0(x, \mathbf{z}_1, s) \to \left(F_1(x, \mathbf{z}_2^1, s), F_1(x, \mathbf{z}_2^2, s), \ldots, F_1(x, \mathbf{z}_2^n, s), Z_1(x, s)\right),$$

or equivalently

$$M^0 : x \to \left(g_s\left(nf_{X_1}^{\mathbf{z}_1}(x)\right), g_s\left(nf_{X_1}^{\mathbf{z}_1}(x) - 1\right), \ldots, g_s\left(nf_{X_1}^{\mathbf{z}_1}(x) - (n-1)\right), f_{X_1}^{\mathbf{z}_1}(x)\right).$$

The networks $\Psi_{i,s}^{\mathbf{z}_1}$ realizing the components $g_s\left(nf_{X_1}^{\mathbf{z}_1}(x) - i\right), i \in [0 : (n-1)]$, of the mapping M^0 all have depth $s + 3$, whereas the network $\Phi^{\mathbf{z}_1}$ implementing the last component of M^0, $f_{X_1}^{\mathbf{z}_1}(x)$, has depth 2. As we want to apply [9, Lemma II.5], we hence need to augment $\Phi^{\mathbf{z}_1}$ to depth $s + 3$. This is effected by exploiting that $\Phi^{\mathbf{z}_1}(x) \geq 0$, $\forall x \in \mathbb{R}$, which allows us to retain the input-output relation realized by the network while amending it by multiplications by 1 (acting as affine transformations) interlaced by applications of ρ for an overall depth of $s + 3$. This leads to the augmented network $\tilde{\Phi}^{\mathbf{z}_1} = \rho \circ \ldots \circ \rho \circ \Phi^{\mathbf{z}_1}$, with $\mathcal{M}(\tilde{\Phi}^{\mathbf{z}_1}) \leq 4n + s - 1$, $\mathcal{L}(\tilde{\Phi}^{\mathbf{z}_1}) = s + 3$. Application of [9, Lemma II.5] now allows us to conclude that the network $\Psi^{M^0} = \left(\Psi_{0,s}^{\mathbf{z}_1}, \Psi_{1,s}^{\mathbf{z}_1}, \ldots, \Psi_{n-1,s}^{\mathbf{z}_1}, \tilde{\Phi}^{\mathbf{z}_1}\right)$ realizing the map M^0 is in $\mathcal{N}_{1,1}$ and satisfies $\mathcal{M}(\Psi^{M^0}) = \mathcal{O}(n^2 + sn)$, $\mathcal{L}(\Psi^{M^0}) = s + 3$. This proves the statement for $r = 0$.

We proceed to the proof for the case $r \geq 1$. To this end, we use (9) to write the map $M^r : \mathbb{R}^{n^r+r} \to \mathbb{R}^{n^{r+1}+r+1}$, for $r \in [1:(d-1)]$, as follows

$$M^r : (y_1, y_2, \ldots, y_{n^r+r}) \to \left(\left[g_s\left(nf_{X_{r+1}}^{\mathbf{z}_{r+1}^i}(y_i) - k\right)\right]_{(i,k) \in ([1:n^r], [0:(n-1)])}, y_{n^r+1}, \ldots, y_{n^r+r}, \sum_{i \in [1:n^r]} f_{X_{r+1}}^{\mathbf{z}_{r+1}^i}(y_i)\right),$$

where the notation $[h(i,k)]_{(i,k) \in ([1:n^r], [0:(n-1)])}$ designates the sequence $h(i,k)$ with (i,k) ranging over $([1:n^r], [0: (n-1)])$, with ordering according to $\left([h(1,k)]_{k \in [0:(n-1)]}, [h(2,k)]_{k \in [0:(n-1)]}, \ldots, [h(n^r,k)]_{k \in [0:(n-1)]}\right)$. As discussed above, each $g_s\left(nf_{X_{r+1}}^{\mathbf{z}_{r+1}^i}(y_i) - k\right)$ can be realized by a network $\Psi_{k,s}^{\mathbf{z}_{r+1}^i} \in \mathcal{N}_{1,1}$ with $\mathcal{M}(\Psi_{k,s}^{\mathbf{z}_{r+1}^i}) \leq 8n + 22s - 8$, $\mathcal{L}(\Psi_{k,s}^{\mathbf{z}_{r+1}^i}) = s + 3$. We will also need the identity networks $\Phi_{id}^{s+3}(x) = (\rho \circ \cdots \circ \rho)(x) = x$, for all $x \geq 0$, with $\mathcal{M}(\Phi_{id}^{s+3}) = s + 3$, $\mathcal{L}(\Phi_{id}^{s+3}) = s + 3$. Finally, by [9, Lemma II.6], there exists a network Ψ^Σ realizing the function $\sum_{i \in [1:n^r]} f_{X_{r+1}}^{\mathbf{z}_{r+1}^i}(y_i)$, and with $\Psi^\Sigma \in \mathcal{N}_{n^r,1}$, $\mathcal{M}(\Psi^\Sigma) \leq 4n^{r+1}$, $\mathcal{L}(\Psi^\Sigma) = 2$. We shall also need the extension of Ψ^Σ to a network of depth $s + 3$ according to $\tilde{\Psi}^\Sigma = \rho \circ \ldots \circ \rho \circ \Psi^\Sigma$ with $\mathcal{M}(\tilde{\Psi}^\Sigma) \leq 4n^{r+1} + s + 1$, $\mathcal{L}(\tilde{\Psi}^\Sigma) = s + 3$. The proof is now concluded by realizing the map M^r as a ReLU network Ψ^{M^r} according to

$$\Psi^{M^r}(y_1, y_2, \ldots, y_{n^r+r})$$
$$= \left(\left[\Psi_{k,s}^{\mathbf{z}_{r+1}^i}(y_i)\right]_{(i,k) \in ([1:n^r], [0:(n-1)])}, \Phi_{id}^{s+3}(y_{n^r+1}), \ldots, \Phi_{id}^{s+3}(y_{n^r+r}), \tilde{\Psi}^\Sigma(y_1, \ldots, y_{n^r})\right).$$

[3]e.g., for $n = 2, r = 3$, the order is $\mathbf{z}_3^1 = (0,0), \mathbf{z}_3^2 = (0,1), \mathbf{z}_3^3 = (1,0), \mathbf{z}_3^4 = (1,1)$.

Application of [9, Lemma II.5] now yields $\Psi^{M^r} \in \mathcal{N}_{n^r+r, n^{r+1}+r+1}$ with $\mathcal{M}(\Psi^{M^r}) = \mathcal{O}(n^{r+2}+sn^{r+1})$, $\mathcal{L}(\Psi^{M^r}) = s + 3$. $\qquad \square$

The next result characterizes the ReLU networks realizing the transport map and quantifies their size in terms of connectivity and depth.

Lemma VI.5. *For every $p_{\mathbf{X}} \in \tilde{\mathcal{E}}_\delta[0,1]_n^d$ with $d > 1$, the corresponding transport map*

$$M : x \rightarrow (Z_1(x,s), Z_2(x,s), \dots, Z_d(x,s))$$

can be realized through a Δ-quantized ReLU network $\Psi^M \in \mathcal{N}_{1,d}$ with $\mathcal{M}(\Psi^M) = \mathcal{O}(n^d + sn^{d-1})$, $\mathcal{L}(\Psi^M) = (s+3)d - s - 1$, and $\Delta = \frac{\delta}{n}$.

Proof. Consider the map $M' := M^{d-2} \circ M^{d-1} \circ \cdots \circ M^0$,

$$M' : x \rightarrow \Big(F_{d-1}(x, \mathbf{z}_d^1, s), F_{d-1}(x, \mathbf{z}_d^2, s), \dots, F_{d-1}(x, \mathbf{z}_d^{n^{d-1}}, s), Z_1(x,s), Z_2(x,s), \dots, Z_{d-1}(x,s) \Big),$$

where the M^r, $r \in [0 : (d-2)]$, are as defined in Lemma VI.4. Since by Lemma VI.4, $M^r, r \in [0 : (d-2)]$, can be realized by a network with connectivity $\mathcal{O}(n^{r+2} + sn^{r+1})$ and depth $s + 3$, it follows from [9, Lemma II.3] that the map M' can be implemented by a network $\Psi' \in \mathcal{N}_{1,n^{d-1}+d-1}$, with $\mathcal{M}(\Psi') = \mathcal{O}(n^d + sn^{d-1})$, $\mathcal{L}(\Psi') = (s+3)(d-1)$; here, we used $\sum_{k=0}^{d-2} \mathcal{O}(n^{k+2} + sn^{k+1}) = \mathcal{O}(n^d + sn^{d-1})$. Next, consider the map

$$S : \Big(y_1, \dots, y_{n^{d-1}+d-1} \Big) \rightarrow \Big(\rho(y_{n^{d-1}+1}), \rho(y_{n^{d-1}+2}), \dots, \rho(y_{n^{d-1}+d-1}), \rho\big(\sum_{i \in [1:n^{d-1}]} y_i \big) \Big),$$

and note that by [9, Lemma II.5], there exists a network $\Psi^S \in \mathcal{N}_{n^{d-1}+d-1,d}$ with $\mathcal{M}(\Psi^S) \leq n^{d-1} + 2d - 1$, and $\mathcal{L}(\Psi^S) = 2$ realizing S. The proof is concluded by noting that, thanks to [9, Lemma II.3], the desired map $M = S \circ M'$ is realized by the network $\Psi^M := \Psi^S(\Psi'(x))$, $\Psi^M \in \mathcal{N}_{1,d}$ with $\mathcal{M}(\Psi^M) = \mathcal{O}(n^d + sn^{d-1})$, $\mathcal{L}(\Psi') = (s+3)d - s - 1$. Moreover, the weights of Ψ^M are either of Type 1 or Type 2 w.r.t. Δ-quantization. $\qquad \square$

We are now ready to state the main result of this section, namely that for every quantized histogram distribution $p_{\mathbf{X}}$ and every $\varepsilon > 0$, there exists a quantized ReLU network Ψ satisfying $W(\Psi \# U, p_{\mathbf{X}}) \leq \varepsilon$. In particular, we also quantify the dependence of ε on the resolution n and the dimension d of $p_{\mathbf{X}}$ as well as the depth of the network Ψ.

Theorem VI.6. *For every δ-quantized $p_{\mathbf{X}} \in \tilde{\mathcal{E}}_\delta[0,1]_n^d$ with $d > 1$, there exists a Δ-quantized ReLU network $\Psi \in \mathcal{N}_{1,d}$ with $\mathcal{M}(\Psi) = \mathcal{O}(n^d + sn^{d-1})$, $\mathcal{L}(\Psi) = (s+3)d - s - 1$, and $\Delta = \frac{\delta}{n}$, such that*

$$W(\Psi \# U, p_{\mathbf{X}}) \leq \frac{\sqrt{d}}{n2^s}. \tag{24}$$

Proof. By Lemma VI.5, for every $p_{\mathbf{X}} \in \tilde{\mathcal{E}}_\delta[0,1]_n^d$ with $d > 1$, the corresponding transport map

$$M : x \rightarrow (Z_1(x,s), Z_2(x,s), \dots, Z_d(x,s))$$

can be realized through a Δ-quantized ReLU network $\Psi^M \in \mathcal{N}_{1,d}$ with $\mathcal{M}(\Psi^M) = \mathcal{O}(n^d + sn^{d-1})$, $\mathcal{L}(\Psi^M) = (s+3)d - s - 1$, and $\Delta = \frac{\delta}{n}$. Moreover, as $\tilde{\mathcal{E}}_\delta[0,1]_n^d \subset \mathcal{E}[0,1]_n^d$, it follows from Theorem V.6 that

$$W(\Psi^M \# U, p_{\mathbf{X}}) \leq \frac{\sqrt{d}}{n2^s}. \qquad \square$$

We note that for fixed histogram resolution n, the upper bound on the approximation error (24) decays exponentially in s and hence in network depth $\mathcal{L}(\Psi)$. In particular, choosing $s \sim n$, guarantees that the error in Theorem VI.6 decays exponentially in n while the connectivity of the network is in $\mathcal{O}(n^d)$; this behavior is asymptotically optimal as the number of parameters in $\tilde{\mathcal{E}}_\delta[0,1]_n^d$ is of the same order.

VII. Approximation of arbitrary distributions on $[0,1]^d$ by quantized histogram distributions

This section is concerned with the approximation of arbitrary distributions ν supported on $[0,1]^d$ by δ-quantized histogram distributions of resolution n as defined in the previous section.

Define the k-dimensional subcube $c_{\mathbf{i}_k} = [i_1/n, (i_1+1)/n] \times [i_2/n, (i_2+1)/n] \times \cdots \times [i_k/n, (i_k+1)/n]$, where $\mathbf{i}_k = (i_1, i_2, \ldots, i_k) \in [0{:}(n-1)]^k$, and its corner point

$$p_{\mathbf{i}_k} = \left(\frac{i_1}{n}, \frac{i_2}{n}, \cdots, \frac{i_k}{n} \right).$$

Next, we discretize the domain $[0,1]^d$ into the subcubes $c_{\mathbf{i}_d}$ and characterize the amount of probability mass ν assigns to the individual subcubes. First, set

$$m_{\mathbf{i}_d} := \nu(c_{\mathbf{i}_d}).$$

Then, for $k \in [1{:}(d-1)]$, we define the projections $P_k : \mathbb{R}^d \to \mathbb{R}^k$, $(x_1, \ldots, x_k, \ldots, x_d) \mapsto (x_1, \ldots, x_k)$ and the corresponding k-dimensional marginals $\nu_k := P_k \# \nu$ with weights

$$m_{\mathbf{i}_k} := \nu_k(c_{\mathbf{i}_k}).$$

It will also be useful to define conditional masses according to $n_{i_1} = m_{i_1}$ and, for $k \in [2{:}d]$, for all[4] $\mathbf{i}_{k-1}$ with $m_{\mathbf{i}_{k-1}} \neq 0$,

$$n_{\mathbf{i}_k} := \frac{m_{\mathbf{i}_k}}{m_{\mathbf{i}_{k-1}}}.$$

For $m_{\mathbf{i}_{k-1}} = 0$, we can, in principle, set the conditional masses arbitrarily, but, for concreteness, we choose

$$n_{\mathbf{i}_k} := \frac{1}{n}.$$

Now that we have defined the masses $m_{\mathbf{i}_k}$ and the conditional masses $n_{\mathbf{i}_k}$ for the distribution ν, we can proceed to derive the masses $\tilde{m}_{\mathbf{i}_k}$ and $\tilde{n}_{\mathbf{i}_k}$ of the corresponding δ-quantized histogram distribution. Denote the index of the subcube with the highest (original) mass in the first coordinate as[5]

$$i_1^{*(\mathbf{i}_0)} := \arg\max_{i_1 \in [0{:}(n-1)]} m_{i_1}. \tag{25}$$

If there are multiple subcubes with the same maximal mass, simply pick one of them (it does not matter which one). Now, for $k = 1$ and $i_1 \neq i_1^{*(\mathbf{i}_0)}$, we choose the quantized masses as follows,

$$\tilde{m}_{i_1} := \tilde{n}_{i_1} := \begin{cases} \delta \lceil \frac{1}{\delta} m_{i_1} \rceil, & \text{if } m_{i_1} > 0 \\ \delta, & \text{if } m_{i_1} = 0 \end{cases},$$

and for $i_1 = i_1^{*(\mathbf{i}_0)}$,

$$\tilde{m}_{i_1^{*(\mathbf{i}_0)}} := \tilde{n}_{i_1^{*(\mathbf{i}_0)}} := 1 - \sum_{i_1 \neq i_1^{*(\mathbf{i}_0)}} \tilde{m}_{i_1}.$$

Note that with this definition, the quantized masses $\tilde{m}_{i_1}$ are always nonzero for $i_1 \neq i_1^{*(\mathbf{i}_0)}$, even in subcubes where the original masses m_{i_1} are equal to zero. We will later verify that this is also the case for $i_1 = i_1^{*(\mathbf{i}_0)}$ whenever $\delta < \frac{1}{n(n-1)}$. For $k \geq 2$, we similarly borrow mass from the subcube with maximum mass, and we do so in each coordinate individually. To this end, for each $k \in [2{:}d]$, we set for all $\mathbf{i}_{k-1}$,

$$i_k^{*(\mathbf{i}_{k-1})} := \arg\max_{i_k \in [0{:}(n-1)]} m_{\mathbf{i}_k}.$$

---

[4] Throughout, we use the symbols $\mathbf{i}_1$ and i_1 interchangeably.

[5] Formally, $\mathbf{i}_0$, albeit not defined, would correspond to a 0-dimensional quantity. It is used throughout the paper only for notational consistency.

As in the assignment (25) for the first coordinate, if there are multiple such values, any of them will do. To define the quantized conditional masses, we set for each $\mathbf{i}_{k-1} \in [0:(n-1)]^{k-1}$ and each $i_k \neq i_k^{*(\mathbf{i}_{k-1})}$,

$$\tilde{n}_{\mathbf{i}_k} := \begin{cases} \delta \lceil \frac{1}{\delta} n_{\mathbf{i}_k} \rceil = \delta \lceil \frac{1}{\delta} \frac{m_{\mathbf{i}_k}}{m_{\mathbf{i}_{k-1}}} \rceil, & \text{if } m_{\mathbf{i}_k} > 0 \\ \delta, & \text{if } m_{\mathbf{i}_k} = 0 \end{cases},$$

as long as $m_{\mathbf{i}_{k-1}} > 0$. If $m_{\mathbf{i}_{k-1}} = 0$, we let

$$\tilde{n}_{\mathbf{i}_k} := \delta \left\lceil \frac{1}{\delta} n_{\mathbf{i}_k} \right\rceil = \delta \left\lceil \frac{1}{\delta} \frac{1}{n} \right\rceil.$$

We can then define the quantized weights according to

$$\tilde{m}_{\mathbf{i}_k} := \tilde{m}_{\mathbf{i}_{k-1}} \tilde{n}_{\mathbf{i}_k} = \tilde{n}_{\mathbf{i}_k} \cdots \tilde{n}_{\mathbf{i}_1}.$$

Finally, for $i_k = i_k^{*(\mathbf{i}_{k-1})}$, we set

$$\tilde{n}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})} \right)} := 1 - \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \tilde{n}_{(\mathbf{i}_{k-1}, i_k)}$$

and correspondingly

$$\tilde{m}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})} \right)} := \tilde{m}_{\mathbf{i}_{k-1}} \tilde{n}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})} \right)} = \tilde{n}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})} \right)} \cdots \tilde{n}_{\mathbf{i}_1}.$$

We now check that the quantized weights verify the following properties:

1) Correct marginals:

$$\sum_{i_k=1}^{n} \tilde{m}_{(\mathbf{i}_{k-1}, i_k)} = \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \tilde{m}_{(\mathbf{i}_{k-1}, i_k)} + \tilde{m}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})} \right)}$$

$$= \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \tilde{m}_{\mathbf{i}_{k-1}} \tilde{n}_{(\mathbf{i}_{k-1}, i_k)} + \tilde{m}_{\mathbf{i}_{k-1}} \tilde{n}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})} \right)}$$

$$= \tilde{m}_{\mathbf{i}_{k-1}} \left(\sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \tilde{n}_{(\mathbf{i}_{k-1}, i_k)} + \left(1 - \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \tilde{n}_{(\mathbf{i}_{k-1}, i_k)} \right) \right)$$

$$= \tilde{m}_{\mathbf{i}_{k-1}}.$$

2) If $\delta < \frac{1}{n(n-1)}$, then all quantized masses are positive. To this end, we first note that

$$n_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})} \right)} = \frac{m_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})} \right)}}{m_{\mathbf{i}_{k-1}}} \geq \frac{1}{n}.$$

Since for $i_k \neq i_k^{*(\mathbf{i}_{k-1})}$, we have by definition

$$\tilde{n}_{\mathbf{i}_k} - n_{\mathbf{i}_k} \leq \delta,$$

it follows that

$$\tilde{n}_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})} \right)} = 1 - \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \tilde{n}_{(\mathbf{i}_{k-1}, i_k)}$$

$$\geq 1 - \sum_{i_k \neq i_k^{*(\mathbf{i}_{k-1})}} \left(n_{(\mathbf{i}_{k-1}, i_k)} + \delta \right)$$

$$= n_{\left(\mathbf{i}_{k-1}, i_k^{*(\mathbf{i}_{k-1})} \right)} - (n-1)\delta$$

$$> \frac{1}{n} - \frac{n-1}{n(n-1)} = 0.$$

We next formalize the procedure for going from the original masses $m_{\mathbf{i}_k}$ to the quantized masses $\tilde{m}_{\mathbf{i}_k}$ by characterizing a transport map effecting this transition.

Lemma VII.1. *Let $k \in [1:d]$, ν a distribution supported on $[0,1]^k$ and with masses $m_{\mathbf{i}_k}$ in the subcubes $c_{\mathbf{i}_k}$ and conditional masses $n_{\mathbf{i}_k}$, all as specified above. Let the quantized masses $\tilde{m}_{\mathbf{i}_k}$ and the conditional quantized masses $\tilde{n}_{\mathbf{i}_k}$ also be given as above. Then, for all $\mathbf{i}_k$, we have*

$$
\tilde{m}_{\mathbf{i}_k} = m_{\mathbf{i}_k} + \sum_{k'=1}^{k} \chi_{[0:(n-1)]\setminus\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'})\, \tilde{\Upsilon}(\mathbf{i}, k, k'+1)\big(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big)\Upsilon^{\eta}(\mathbf{i}, k'-1, 1)
$$

$$
- \sum_{k'=1}^{k} \chi_{\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'})\Upsilon(\mathbf{i}, k, k'+1)\big(n_{\mathbf{i}_{k'}} - \tilde{n}_{\mathbf{i}_{k'}}\big)\Upsilon^{\eta}(\mathbf{i}, k'-1, 1),
$$

(26)

where

$$
\Upsilon(\mathbf{i}, b, a) = \begin{cases} n_{\mathbf{i}_b} \cdots n_{\mathbf{i}_a}, & \text{if } b \geq a \\ 1, & \text{else} \end{cases},
$$

$$
\tilde{\Upsilon}(\mathbf{i}, b, a) = \begin{cases} \tilde{n}_{\mathbf{i}_b} \cdots \tilde{n}_{\mathbf{i}_a}, & \text{if } b \geq a \\ 1, & \text{else} \end{cases},
$$

and

$$
\Upsilon^{\eta}(\mathbf{i}, b, a) = \begin{cases} \eta_{\mathbf{i}_b}(i_b) \cdots \eta_{\mathbf{i}_a}(i_a), & \text{if } b \geq a \\ 1, & \text{else} \end{cases},
$$

with

$$
\eta_{\mathbf{i}_k}(i_k) := \begin{cases} n_{\mathbf{i}_k}, & \text{if } i_k \neq i_k^{*(\mathbf{i}_{k-1})} \\ \tilde{n}_{\mathbf{i}_k}, & \text{if } i_k = i_k^{*(\mathbf{i}_{k-1})} \end{cases}.
$$

The proof of Lemma VII.1 is provided in the appendix.

We are now ready to state the main result of this section. Specifically, we establish an upper bound on the Wasserstein distance between a given (arbitrary) distribution ν supported on $[0,1]^d$, for any $d \in \mathbb{N}$, and the corresponding δ-quantized histogram distribution of resolution n obtained based on the procedure described above.

Theorem VII.2. *Let $d \in \mathbb{N}$. For every distribution ν supported on $[0,1]^d$, there exists a δ-quantized histogram distribution μ of resolution n such that*

$$
W(\mu, \nu) \leq \frac{2\sqrt{d}}{n} + \frac{d(d+1)}{2}(n-1)\delta.
$$

Proof. The proof proceeds in three steps as follows.

1) For each $\mathbf{i}_d \in [0:(n-1)]^d$, we redistribute the mass $m_{\mathbf{i}_d} = \nu(c_{\mathbf{i}_d})$ to a point mass concentrated in the corner point $p_{\mathbf{i}_d}$.

2) We transport the masses $m_{\mathbf{i}_d}$ according to the procedure described above to result in the masses $\tilde{m}_{\mathbf{i}_d}$, still located at $p_{\mathbf{i}_d}$.

3) For each $\mathbf{i}_d \in [0:(n-1)]^d$, we spread out the mass $\tilde{m}_{\mathbf{i}_d}$ uniformly across the subcube indexed by $\mathbf{i}_d$.

Step 1. We define the distribution

$$
\nu' = \sum_{\mathbf{i}_d \in [0:(n-1)]^d} m_{\mathbf{i}_d} \delta_{p_{\mathbf{i}_d}}
$$

24

and note that transporting ν to ν' incurs transportation cost

$$W(\nu, \nu') \leq \sum_{\mathbf{i}_d} m_{\mathbf{i}_d} \underbrace{\sqrt{\left(\frac{1}{n}\right)^2 + \cdots + \left(\frac{1}{n}\right)^2}}_{\text{maximum distance in each subcube}} = \sum_{\mathbf{i}_d} m_{\mathbf{i}_d} \frac{\sqrt{d}}{n} = \frac{\sqrt{d}}{n}. \tag{27}$$

Step 2. To redistribute the masses from the original values $m_{\mathbf{i}_d}$ to the quantized values $\tilde{m}_{\mathbf{i}_d}$, we proceed co-ordinate by coordinate. Specifically, in the k-th coordinate, we carry out two (sets of) transportations. The first one moves, for fixed $i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_d$, mass from the point $p_{(i_1, \ldots, i_{k-1}, i_k^{*(\mathbf{i}_{k-1})}, i_{k+1}, \ldots, i_d)}$ to the points $p_{(i_1, \ldots, i_{k-1}, i_k, i_{k+1}, \ldots, i_d)}$, for all $i_k \neq i_k^{*(\mathbf{i}_{k-1})}$, and does this for all tuples $i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_d$. The second set of transportations reconfigures masses in the coordinates $[1:(k-1)]$ so as to obtain the correct marginals in coordinate k. These reconfigurations moreover preserve the marginals in coordinates $[1:(k-1)]$. We make all this precise through the following claim, proved below after Step 3 has been presented.

Claim: Reconfiguring the masses between the corner points such that the mass in the point $p_{\mathbf{i}_d}$ is given by

$$m_{\mathbf{i}_d} + \sum_{k'=1}^{k} \chi_{[0:(n-1)]\setminus\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \left(\frac{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}, i_{k+1}, \ldots, i_d\right)}}{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}}\right)$$
$$\tilde{\Upsilon}(\mathbf{i}, k, k'+1)\left(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\right) \Upsilon^{\eta}(\mathbf{i}, k'-1, 1) \tag{28}$$
$$- \sum_{k'=1}^{k} \chi_{\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \, \Upsilon(\mathbf{i}, d, k'+1)\left(n_{\mathbf{i}_{k'}} - \tilde{n}_{\mathbf{i}_{k'}}\right) \Upsilon^{\eta}(\mathbf{i}, k'-1, 1),$$

where

$$m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}, i_{k+1}, \ldots, i_d\right)} := \sum_{i_{k'+1}, \ldots, i_k} m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}, i_{k'+1}, \ldots, i_d\right)},$$

yields the correct marginal masses $\tilde{m}_{\mathbf{i}_{k'}}$ in all coordinates $k' \in [1:k]$ and comes at a Wasserstein cost of at most $k(n-1)\delta$, i.e., the Wasserstein distance between the configuration of masses before the moves and the configuration after the moves is at most $k(n-1)\delta$. There is a slight complication when $m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}\right)} = 0$ as in this case the fraction in (28) is technically undefined. However, analogously to the definition of the $n_{\mathbf{i}_k}$ in the case of zero-masses in the discussion preceding this theorem, we take

$$\left(\frac{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}, i_{k+1}, \ldots, i_d\right)}}{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}}\right) := \left(\frac{1}{n}\right)^{d-k}$$

when $m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}\right)} = 0$. In either case, we have

$$\sum_{i_{k+1}, \ldots, i_d} \left(\frac{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}, i_{k+1}, \ldots, i_d\right)}}{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}}\right) = 1.$$

We note that the transport map in the Claim characterizes, at a high level, the state of the masses at an intermediate step in the transportation, while (26) describes the "final state" after all the moves have been completed in coordinate k.

If we accept the claim and apply it for $k = d$ in combination with Lemma VII.1, it follows that the masses $m_{\mathbf{i}_d}$ are, indeed, redistributed to the masses $\tilde{m}_{\mathbf{i}_d}$. Moreover, we get that the total cost of the transportations in Step 2 effecting this redistribution is upper-bounded by

$$(n-1)\,\delta + 2\,(n-1)\,\delta + \cdots + d\,(n-1)\,\delta = \frac{d(d+1)}{2}(n-1)\,\delta.$$

Step 3. The Wasserstein cost associated with spreading out the masses $\tilde{m}_{\mathbf{i}_d}$ uniformly across their associated subcubes follows from (27) as

$$W\left(\sum_{\mathbf{i}_d} \tilde{m}_{\mathbf{i}_d}\delta_{p_{\mathbf{i}_d}}, \mu\right) \leq \frac{\sqrt{d}}{n}.$$

Using the fact that Wasserstein distance is a metric, we can put the costs incurred in the individual steps together according to

$$W(\mu,\nu) \leq \underbrace{W\left(\nu, \sum_{\mathbf{i}_d} m_{\mathbf{i}_d}\delta_{p_{\mathbf{i}_d}}\right)}_{\text{Step 1}} + \underbrace{W\left(\sum_{\mathbf{i}_d} m_{\mathbf{i}_d}\delta_{p_{\mathbf{i}_d}}, \sum_{\mathbf{i}_d} \tilde{m}_{\mathbf{i}_d}\delta_{p_{\mathbf{i}_d}}\right)}_{\text{Step 2}} + \underbrace{W\left(\sum_{\mathbf{i}_d} \tilde{m}_{\mathbf{i}_d}\delta_{p_{\mathbf{i}_d}}, \mu\right)}_{\text{Step 3}}$$

$$\leq \frac{\sqrt{d}}{n} + \frac{d(d+1)}{2}(n-1)\delta + \frac{\sqrt{d}}{n}$$

$$= \frac{2\sqrt{d}}{n} + \frac{d(d+1)}{2}(n-1)\delta.$$

It remains to prove the claim.

Proof of the Claim. We proceed by induction on k and start with the base case $k = 1$. The statement on the transportation cost associated with (28) does not need an induction argument, rather it follows as a byproduct of the proof by induction. Evaluating the transport map for $k = 1$ yields

$$m_{\mathbf{i}_d} + \chi_{[0:(n-1)]\setminus\{i_1^{*(\mathbf{i}_0)}\}}(i_1)\left(\frac{m_{\left(i_1^{*(\mathbf{i}_0)},i_2,\ldots,i_d\right)}}{m_{i_1^{*(\mathbf{i}_0)}}}\right)(\tilde{m}_{i_1} - m_{i_1})$$

$$- \chi_{\{i_1^{*(\mathbf{i}_0)}\}}(i_1)\left(\frac{m_{\left(i_1^{*(\mathbf{i}_0)},i_2,\ldots,i_d\right)}}{m_{i_1^{*(\mathbf{i}_0)}}}\right)(m_{i_1} - \tilde{m}_{i_1}).$$

Since masses are moved in the first coordinate only and $(\tilde{m}_{i_1} - m_{i_1}) \leq \delta$, for $i_1 \neq i_1^{*(\mathbf{i}_0)}$, the Wasserstein cost of the overall transportation satisfies

$$\sum_{i_1 \neq i_1^{*(\mathbf{i}_0)}} \sum_{i_2,\ldots,i_d} \left(\frac{m_{\left(i_1^{*(\mathbf{i}_0)},i_2,\ldots,i_d\right)}}{m_{i_1^{*(\mathbf{i}_0)}}}\right)(\tilde{m}_{i_1} - m_{i_1}) \leq (n-1)\delta.$$

Furthermore, we obtain the desired marginal masses in i_1 as a consequence of

$$\sum_{i_2,\ldots,i_d}\left(m_{\mathbf{i}_d} + \left(\frac{m_{\left(i_1^{*(\mathbf{i}_0)},i_2,\ldots,i_d\right)}}{m_{i_1^{*(\mathbf{i}_0)}}}\right)(\tilde{m}_{i_1} - m_{i_1})\right) = m_{i_1} + (\tilde{m}_{i_1} - m_{i_1}) = \tilde{m}_{i_1}.$$

This completes the proof of the base case.

We proceed to establish the induction step. Assume that transportations were conducted in coordinate k according to (28) and that all marginal masses up to and including coordinate k are as desired. We consider the transport equation (28) in coordinate $k+1$, i.e., the sums in (28) range from 1 to $k+1$ and start by pointing out that

$$n_{\left(i_1,\ldots,i_{k+1}^{*(\mathbf{i}_k)},\ldots,i_d\right)}n_{\left(i_1,\ldots,i_{k+1}^{*(\mathbf{i}_k)},\ldots,i_{d-1}\right)}\cdots n_{\left(i_1,\ldots,i_{k+1}^{*(\mathbf{i}_k)},i_{k+2}\right)} = \left(\frac{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)}\right)}}\right).$$

The first set of transportations (corresponding to the index $k' = k+1$ in the transport equation (28) evaluated for coordinate $k+1$) hence amounts to moving, for fixed $i_1,\ldots,i_k,i_{k+2},\ldots,i_d$, the mass

$$\left(\frac{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)}\right)}}\right)(n_{\mathbf{i}_{k+1}} - \tilde{n}_{\mathbf{i}_{k+1}})\Upsilon^\eta(\mathbf{i},k,1)$$

out of the point $p_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)},i_{k+2},\ldots,i_d\right)}$ and redistributing it across the points $p_{(i_1,\ldots,i_k,i_{k+1},i_{k+2},\ldots,i_d)}$, for $i_{k+1} \neq i_{k+1}^{*(\mathbf{i}_k)}$. Note that for $i_{k+1} = i_{k+1}^{*(\mathbf{i}_k)}$, the quantity $\left(n_{\mathbf{i}_{k+1}} - \tilde{n}_{\mathbf{i}_{k+1}}\right)$ is positive by definition of $\tilde{n}$. These transportations are conducted for all possible tuples $i_1,\ldots,i_k,i_{k+2},\ldots,i_d$. The Wasserstein cost associated with the collection of these transportations satisfies

$$
\sum_{i_1,\ldots,i_k} \sum_{i_{k+1}\neq i_{k+1}^{*(\mathbf{i}_k)}} \sum_{i_{k+2},\ldots,i_d} \left(\frac{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_k,i_{k+1}^{*(\mathbf{i}_k)}\right)}}\right) \left(\tilde{n}_{\mathbf{i}_{k+1}} - n_{\mathbf{i}_{k+1}}\right)\Upsilon^\eta(\mathbf{i},k,1)
$$

$$
= \sum_{i_1,\ldots,i_k} \sum_{i_{k+1}\neq i_{k+1}^{*(\mathbf{i}_k)}} \left(\tilde{n}_{\mathbf{i}_{k+1}} - n_{\mathbf{i}_{k+1}}\right)\Upsilon^\eta(\mathbf{i},k,1)
$$

$$
\leq (n-1)\,\delta \sum_{i_1,\ldots,i_k} \Upsilon^\eta(\mathbf{i},k,1)
$$

$$
\leq (n-1)\,\delta \sum_{i_1,\ldots,i_k} \Upsilon(\mathbf{i},k,1)
$$

$$
= (n-1)\,\delta,
$$

where the last inequality follows because $\eta_{\mathbf{i}_k}(i_k) = \tilde{n}_{\mathbf{i}_k}$ exactly when $i_k = i_k^{*(\mathbf{i}_{k-1})}$, in which case we have $\tilde{n}_{\mathbf{i}_k} \leq n_{\mathbf{i}_k}$. The second set of transportations reconfigures the masses in the coordinates $k' \leq k$ in order to obtain correct marginals in the $(k+1)$-th coordinate. To this end, we first note that, for each $k' \leq k$, for all $i_{k'}$, the following identity holds

$$
\left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+1},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)}\right)}}\right) = \underbrace{\left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)}\right)}}\right)}_{\text{independent of } i_{k+1}} \underbrace{\left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+1},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+2},\ldots,i_d\right)}}\right)}_{\text{sums to 1 in } i_{k+1}},
$$

as long as $m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+2},\ldots,i_d\right)} \neq 0$. If $m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+2},\ldots,i_d\right)} = 0$, then, as usual, we set

$$
\left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+1},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+2},\ldots,i_d\right)}}\right) := \frac{1}{n}.
$$

Specifically, noting that by the induction assumption transportation according to (28) was carried out in coordinate k, we would like to reconfigure, for each $k' \leq k$, the masses

$$
\left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)}\right)}}\right)\left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+1},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+2},\ldots,i_d\right)}}\right)\tilde{\Upsilon}(\mathbf{i},k,k'+1)
$$

$$
\left(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\right)\Upsilon^\eta(\mathbf{i},k'-1,1) \tag{29}
$$

into

$$
\left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+2},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)}\right)}}\right)\tilde{n}_{\mathbf{i}_{k+1}}\tilde{\Upsilon}(\mathbf{i},k,k'+1)\left(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\right)\Upsilon^\eta(\mathbf{i},k'-1,1). \tag{30}
$$

This reconfiguration is possible as only one term in each (29) and (30) depends on i_{k+1} and

$$
\sum_{i_{k+1}} \tilde{n}_{\mathbf{i}_{k+1}} = 1 = \sum_{i_{k+1}} \left(\frac{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+1},\ldots,i_d\right)}}{m_{\left(i_1,\ldots,i_{k'}^{*(\mathbf{i}_{k'}-1)},i_{k+2},\ldots,i_d\right)}}\right).
$$

Masses to be moved in this manner appear for all $i_1, ..., i_{k'-1}, i_{k'}, i_{k'+1}, \ldots, i_k, i_{k+2}, \ldots, i_d$ with $i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}$. It follows by inspection of the transport map (28) that these transportations do not alter the marginals for $k' \leq k$ as, for given k', the mass moved out of the point $p_{\left(i_1, \ldots, i_{k'-1}, i_{k'}^{*(\mathbf{i}_{k'-1})}, i_{k'+1}, \ldots, i_d\right)}$, accounted for by the sum with negative sign in (28), equals the total mass moved into the points $p_{(i_1, \ldots, i_{k'-1}, i_{k'}, i_{k'+1}, \ldots, i_d)}$, for $i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}$, accounted for by the sum with positive sign. These moves hence retain the marginals for $k' \leq k$, which are correct by the induction assumption. Before establishing that the desired marginals in coordinate $k+1$ are obtained, we compute the Wasserstein cost associated with the mass reconfiguration moves according to

$$
\sum_{i_1, \ldots, i_{k'-1}} \sum_{i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}} \sum_{i_{k'+1}, \ldots, i_k} \sum_{i_{k+2}, \ldots, i_d} \left(\frac{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}, i_{k+2}, \ldots, i_d\right)}}{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}} \right) \tilde{\Upsilon}(\mathbf{i}, k, k'+1)
$$

$$
\left(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\right) \Upsilon^\eta(\mathbf{i}, k'-1, 1)
$$

$$
= \sum_{i_1, \ldots, i_{k'-1}} \sum_{i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}} \sum_{i_{k'+1}, \ldots, i_k} \tilde{\Upsilon}(\mathbf{i}, k, k'+1)\left(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\right) \Upsilon^\eta(\mathbf{i}, k'-1, 1)
$$

$$
= \sum_{i_1, \ldots, i_{k'-1}} \sum_{i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}} \underbrace{\left(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\right)}_{\leq \delta} \Upsilon^\eta(\mathbf{i}, k'-1, 1)
$$

$$
\leq (n-1)\delta \sum_{i_1, \ldots, i_{k'-1}} \Upsilon^\eta(\mathbf{i}, k'-1, 1)
$$

$$
\leq (n-1)\delta \sum_{i_1, \ldots, i_{k'-1}} \Upsilon(\mathbf{i}, k'-1, 1)
$$

$$
= (n-1)\delta.
$$

We must carry this out for all $k' \in [1:k]$, which results in a Wasserstein cost of $k(n-1)\delta$ for the reconfigurations. Altogether, we have a Wasserstein cost of

$$
(k+1)(n-1)\delta
$$

incurred by the moves corresponding to coordinate $k+1$.

Next, using $\tilde{n}_{\mathbf{i}_{k+1}}\tilde{\Upsilon}(\mathbf{i}, k, k'+1) = \tilde{\Upsilon}(\mathbf{i}, k+1, k'+1)$, it follows that the updated mass in the point $p_{\mathbf{i}_d}$ is given by

$$
m_{\mathbf{i}_d} + \sum_{k'=1}^{k+1} \chi_{[0:(n-1)]\setminus\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \left(\frac{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}, i_{k+2}, \ldots, i_d\right)}}{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}} \right)
$$

$$
\tilde{\Upsilon}(\mathbf{i}, k+1, k'+1)\left(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\right) \Upsilon^\eta(\mathbf{i}, k'-1, 1)
$$

$$
- \sum_{k'=1}^{k+1} \chi_{\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \Upsilon(\mathbf{i}, d, k'+1)\left(n_{\mathbf{i}_{k'}} - \tilde{n}_{\mathbf{i}_{k'}}\right) \Upsilon^\eta(\mathbf{i}, k'-1, 1).
$$

Finally, we need to check that the marginals in the $(k+1)$-th coordinate are, indeed, given by $\tilde{m}_{\mathbf{i}_{k+1}}$. This is accomplished by noting that owing to Lemma VII.1, we have

$$
\sum_{i_{k+2}, \ldots, i_d} \left[m_{\mathbf{i}_d} + \sum_{k'=1}^{k+1} \chi_{[0:(n-1)]\setminus\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \left(\frac{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}, i_{k+2}, \ldots, i_d\right)}}{m_{\left(i_1, \ldots, i_{k'}^{*(\mathbf{i}_{k'-1})}\right)}} \right) \right.
$$

$$
\tilde{\Upsilon}(\mathbf{i}, k+1, k'+1)\left(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\right) \Upsilon^\eta(\mathbf{i}, k'-1, 1)
$$

$$
\left. - \sum_{k'=1}^{k+1} \chi_{\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \Upsilon(\mathbf{i}, d, k'+1)\left(n_{\mathbf{i}_{k'}} - \tilde{n}_{\mathbf{i}_{k'}}\right) \Upsilon^\eta(\mathbf{i}, k'-1, 1) \right]
$$

$$= m_{\mathbf{i}_{k+1}}$$
$$+ \sum_{k'=1}^{k+1} \chi_{[0:(n-1)]\setminus\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \; \tilde{\Upsilon}(\mathbf{i}, k+1, k'+1)\big(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big)\Upsilon^{\eta}(\mathbf{i}, k'-1, 1)$$
$$- \sum_{k'=1}^{k+1} \chi_{\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) \; \Upsilon(\mathbf{i}, k+1, k'+1)\big(n_{\mathbf{i}_{k'}} - \tilde{n}_{\mathbf{i}_{k'}}\big)\Upsilon^{\eta}(\mathbf{i}, k'-1, 1)$$
$$= \tilde{m}_{\mathbf{i}_{k+1}}.$$

This concludes the proof of the induction step and hence establishes the claim. $\qquad\square$

VIII. Approximation of arbitrary distributions on bounded subsets of $\mathbb{R}^d$ with generative ReLU networks

In this section, we put all the pieces developed together and state the main result of this paper.

Theorem VIII.1. *For every distribution ν supported on $[0,1]^d$, there exists a $\frac{\delta}{n}$-quantized ReLU network $\Phi \in \mathcal{N}_{1,d}$ with $\mathcal{M}(\Phi) = \mathcal{O}(n^d + sn^{d-1})$ and $\mathcal{L}(\Phi) = (s+3)d - s - 1$ such that*

$$W(\Phi\#U, \nu) \leq \frac{\sqrt{d}}{n2^s} + \frac{2\sqrt{d}}{n} + \frac{d(d+1)}{2}(n-1)\delta. \tag{31}$$

Proof. The proof follows by application of the triangle inequality for Wasserstein distance in combination with Theorems VI.6 and VII.2 according to

$$W(\Phi\#U, \nu) \leq W(\Phi\#U, \mu) + W(\mu, \nu) \leq \frac{\sqrt{d}}{n2^s} + \frac{2\sqrt{d}}{n} + \frac{d(d+1)}{2}(n-1)\delta,$$

where μ denotes the δ-quantized histogram distribution of resolution n per Theorem VII.2. $\qquad\square$

When the target distribution is uniform, Theorem VIII.1 recovers [2, Theorem 2.1]. We can simplify the bound (31) by setting $\delta = \frac{1}{\lceil\sqrt{d}(d+1)n(n-1)\rceil}$ (which satisfies the requirement $\delta < \frac{1}{n(n-1)}$ guaranteeing that all quantized weights are positive) to obtain, for $n \geq 2$,

$$W(\Phi\#U, \nu) \leq \frac{\sqrt{d}}{n2^s} + \frac{2\sqrt{d}}{n} + \frac{\sqrt{d}}{2n} = \frac{\sqrt{d}}{n2^s} + \frac{5\sqrt{d}}{2n}. \tag{32}$$

The error bound in (32) illustrates the main conceptual insight of this paper, namely that generating arbitrary d-dimensional distributions from a 1-dimensional uniform distribution by pushforward through a deep ReLU network does not come at a cost—in terms of Wasserstein-distance error—relative to generating the target distribution from d independent random variables. Specifically, if we let the depth s of the generating network go to infinity, the first term in the rightmost expression of (32) will go to zero exponentially fast in s—thanks to the space-filling property of the transport map realized by the generating network—leaving us only with the second term, which reflects the error stemming from the histogram approximation of the distribution. Moreover, this second term is inversely proportional to the histogram resolution n and can thus be made arbitrarily small by letting the histogram resolution n approach infinity. The width of the corresponding generating network will grow according to $\mathcal{O}(n^d)$.

Theorem VIII.1 applies to distributions supported on the unit cube $[0,1]^d$. The extension to distributions supported on bounded subsets of $\mathbb{R}^d$ is, however, fairly straightforward. Before stating this extension, we provide a lemma that will help us deal with the scaling and shifting of distributions.

Lemma VIII.2. *Let μ, ν be distributions on $\mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a Lipschitz-continuous mapping with Lipschitz constant $\mathrm{Lip}(f) < \infty$. Then,*

$$W(f\#\mu, f\#\nu) \leq \mathrm{Lip}(f)\, W(\mu, \nu).$$

Proof. Let π be a coupling between μ and ν and let $g : \mathbb{R}^{2d} \to \mathbb{R}^{2d}; (\mathbf{y}_1, \mathbf{y}_2) \mapsto (f(\mathbf{y}_1), f(\mathbf{y}_2))$. Then $g\#\pi$ is a coupling between μ and ν and

$$
\begin{aligned}
W(f\#\mu, f\#\nu) &\leq \int_{\mathbb{R}^{2d}} \|\mathbf{y}_1 - \mathbf{y}_2\| \, d(g\#\pi)(\mathbf{y}_1, \mathbf{y}_2) \\
&= \int_{\mathbb{R}^{2d}} \|f(\mathbf{y}_1) - f(\mathbf{y}_2)\| \, d\pi(\mathbf{y}_1, \mathbf{y}_2) \\
&\leq \text{Lip}(f) \int_{\mathbb{R}^{2d}} \|\mathbf{y}_1 - \mathbf{y}_2\| \, d\pi(\mathbf{y}_1, \mathbf{y}_2) = \text{Lip}(f) \, W(\mu, \nu). \quad \square
\end{aligned}
$$

We are now ready to state the extension announced above.

Theorem VIII.3. *Let ν be a distribution on $\mathbb{R}^d$ supported on $S = \alpha\,[0,1]^d + \boldsymbol{\beta}$ for $\alpha > 0$ and $\boldsymbol{\beta} \in \mathbb{R}^d$. Let $g(\mathbf{x}) = \frac{1}{\alpha}(\mathbf{x} - \boldsymbol{\beta})$ and $\delta = \frac{1}{\lceil \sqrt{d}(d+1)n(n-1) \rceil}$. Then, there exists a $\frac{\delta}{n}$-quantized ReLU network $\Phi \in \mathcal{N}_{1,d}$ with $\mathcal{M}(\Phi) = \mathcal{O}(n^d + sn^{d-1})$ and $\mathcal{L}(\Phi) = (s+3)d - s - 1$ such that*

$$
W(g^{-1}\#\Phi\#U, \nu) \leq \alpha\left(\frac{\sqrt{d}}{n2^s} + \frac{5\sqrt{d}}{2n} \right).
$$

Proof. We first note that g^{-1} is Lipschitz with $\text{Lip}(g^{-1}) = \alpha$. The result then follows immediately from Lemma VIII.2 combined with (32) by taking $\Phi \in \mathcal{N}_{1,d}$ to approximate the distribution $g\#\nu$ according to Theorem VIII.1. $\square$

We finally remark that g^{-1} by virtue of being an affine map can easily be realized by a ReLU network.

IX. COMPLEXITY OF GENERATIVE NETWORKS

In this section, we compare the complexity of ReLU networks generating a given class of probability distributions to fundamental bounds on the complexity of encoding classes of probability distributions through discrete approximations, a process commonly referred to as quantization [12]. Specifically, complexity will be measured in terms of the number of bits needed to describe the generative networks and, respectively, distributions. We begin by reviewing a fundamental result on the approximation of (non-singular) distributions.

Definition IX.1 ([12]). *For $n \in \mathbb{N}$ and the non-singular distribution ν supported on $[0,1]^d$, we define the minimal n-term quantization error as*
$$
V_n(\nu) := \inf\{W(\nu, \mu) : |\text{supp}(\mu)| \leq n\}.
$$

The quantity $V_n(\nu)$ characterizes the approximation error—in Wasserstein distance—incurred by the best discrete n-point approximation of ν.

The next result, taken from [12], states that this approximation error exhibits the same asymptotics for all (non-singular) distributions satisfying a mild moment constraint.

Theorem IX.2 ([12, Theorem 6.2]). *Let $\mathbf{X}$ be a random vector in $\mathbb{R}^d$ with $\mathbf{X} \sim \nu$, where ν is non-singular and supported on $[0,1]^d$, and $\mathbb{E}\|\mathbf{X}\|^{1+\delta} < \infty$ for some $\delta > 0$, where $\|\cdot\|$ is any norm on $\mathbb{R}^d$. Then,*
$$
\lim_{n \to \infty} n^{1/d} V_n(\nu) = C,
$$
where $C > 0$ is a constant depending on d only.

Theorem IX.2 allows us to conclude that the best-approximating discrete distribution must have at least $n = \mathcal{O}(\varepsilon^{-d})$ points for $V_n(\nu) \leq \varepsilon$ to hold. As Wasserstein distance is a metric, we hence have a covering argument which says that the class of (non-singular) distributions ν supported on $[0,1]^d$ (and satisfying the moment constraint in Theorem IX.2) has metric entropy lower-bounded by $d\log(\varepsilon^{-1})$ bits. Although this lower bound is very generous, we demonstrate next that it is achieved for quantized histogram target distributions encoded by their generating ReLU networks.

Lemma IX.3. *Consider the class of quantized histogram distributions $\tilde{\mathcal{E}}_\delta[0,1]_n^d$ and let $\varepsilon \in (0,1/2)$. Then, there exists a set of $\frac{\delta}{n}$-quantized ReLU networks $\Phi(\varepsilon, \cdot)$ of cardinality $2^{\ell(\varepsilon)}$, where $\ell(\varepsilon) \leq C \log(\varepsilon^{-1})$, with C a constant depending on d, δ, n, such that*

$$\sup_{\nu \in \tilde{\mathcal{E}}_\delta[0,1]_n^d} W(\Phi(\varepsilon, \nu)\#U, \nu) \leq \varepsilon.$$

Proof. By Theorem VI.6, for every distribution $\nu \in \tilde{\mathcal{E}}_\delta[0,1]_n^d$, there exists a $\frac{\delta}{n}$-quantized ReLU network Φ, with $\mathcal{M}(\Phi) = \mathcal{O}(n^d + sn^{d-1})$ and $\mathcal{L}(\Phi) = (s+3)d - s - 1$, such that

$$W(\Phi\#U, \nu) \leq \frac{\sqrt{d}}{n2^s} =: \varepsilon.$$

Note that d, n, δ are fixed and ε, as a function of s, can be made arbitrarily small by taking s and hence network depth to be sufficiently large. In particular, network depth needs to scale according to $\mathcal{O}(\log(\varepsilon^{-1}))$. The resulting network Φ will hence depend on ν and ε, indicated by the notation $\Phi(\varepsilon, \nu)$ used henceforth. Next, using $\frac{\delta}{n} \leq 1/2$, which is by assumption, it follows from [9, Proposition VI.7] that the number of bits needed to encode $\Phi(\varepsilon, \nu)$ in a uniquely decodable fashion satisfies

$$\ell(\varepsilon) \leq C_0 \left(\mathcal{M}(\Phi(\varepsilon, \nu)) \log(\mathcal{M}(\Phi(\varepsilon, \nu))) + 1 \right) \log(n/\delta) \leq C(d, \delta, n) \log(\varepsilon^{-1}). \tag{33}$$

$\square$

Remark. We note that the quantized networks considered in the present paper differ slightly from those in [9] as here we employ two types of quantization, namely Type 1 and Type 2 (see Definition VI.3), while in [9] all weights are encoded using Type-1 quantization. This does, however, not have an impact on the bound on $\ell(\varepsilon)$ in (33), in fact, only the constant C_0 changes relative to [9, Proposition VI.7]. More specifically, to encode the quantized weights in the generative networks considered here, we only need one additional bit per weight signifying whether the weight is quantized according to Type 1 or Type 2.

Lemma IX.3 tells us that encoding (or quantizing in the sense of [12]) the class of quantized histogram distributions by pushing forward a scalar uniform distribution through generative ReLU networks achieves the metric entropy limit of $\mathcal{O}(\log(\varepsilon^{-1}))$ as identified in Theorem IX.2. We hasten to add that a metric entropy scaling of $\mathcal{O}(\log(\varepsilon^{-1}))$ for (quantized) histogram distributions of dimension d, resolution n, and quantization level δ, all fixed, is what one would expect as we essentially have to encode polynomially (in n) many (quantized) real numbers. For general (non-singular) distributions, which constitute a much richer class than (quantized) histogram distributions, we can establish an $\mathcal{O}(\varepsilon^{-d})$ (up to a multiplicative log-term) complexity scaling for their corresponding generative networks, formalized as follows.

Lemma IX.4. *Consider the class of non-singular distributions supported on $[0,1]^d$, denoted by $\mathcal{F}([0,1]^d)$, and let $\varepsilon \in (0,1/2)$. Then, there exists a set of quantized ReLU networks $\Phi(\varepsilon, \cdot)$ of cardinality $2^{\ell(\varepsilon)}$, where $\ell(\varepsilon) \leq C\varepsilon^{-d} \log^2(\varepsilon^{-1})$, with C a constant depending on d, such that*

$$\sup_{\nu \in \mathcal{F}([0,1]^d)} W(\Phi(\varepsilon, \nu)\#U, \nu) \leq \varepsilon.$$

Proof. By Theorem VIII.1, for every distribution ν supported on $[0,1]^d$, there exists a $\frac{\delta}{n}$-quantized ReLU network Φ, with $\mathcal{M}(\Phi) = \mathcal{O}(n^d + sn^{d-1})$ and $\mathcal{L}(\Phi) = (s+3)d - s - 1$, such that

$$W(\Phi\#U, \nu) \leq \frac{\sqrt{d}}{n2^s} + \frac{2\sqrt{d}}{n} + \frac{d(d+1)}{2}(n-1)\delta.$$

Setting $\delta = \frac{1}{n^2 d^2}$ and $s = \log(n)$, we hence get

$$W(\Phi\#U, \nu) \leq \frac{3\sqrt{d}+1}{n} =: \varepsilon,$$

for a $\frac{1}{n^3 d^2}$-quantized network $\Phi(\varepsilon, \nu)$, with $\mathcal{M}(\Phi(\varepsilon, \nu)) = \mathcal{O}(n^d)$. Application of [9, Proposition VI.7] allows us to conclude that the number of bits needed to encode $\Phi(\varepsilon, \nu)$ in a uniquely decodable fashion satisfies $\ell(\varepsilon) \leq C_0 \left(\mathcal{M}(\Phi) \log(\mathcal{M}(\Phi)) + 1 \right) \log(n^3 d^2) \leq C(d)\varepsilon^{-d} \log^2(\varepsilon^{-1})$. We note that $C(d)$ scales very unfavorably in d, namely according to $d^{d/2}$. Finally, we remark that the application of [9, Proposition VI.7] requires that $\frac{1}{n^3 d^2} \leq 1/2$, which is satisfied if at least one of n, d is strictly larger than 1. $\square$

31

Proof. Note first that

$$\chi_{[0:(n-1)]\setminus\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) = \begin{cases} 0, & \text{if } i_{k'} = i_{k'}^{*(\mathbf{i}_{k'-1})} \\ 1, & \text{if } i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})} \end{cases}$$

and

$$\chi_{\left\{i_{k'}^{*(\mathbf{i}_{k'-1})}\right\}}(i_{k'}) = \begin{cases} 0, & \text{if } i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})} \\ 1, & \text{if } i_{k'} = i_{k'}^{*(\mathbf{i}_{k'-1})} \end{cases},$$

so for a given $i_{k'}$ only one of the two χ-terms above is active. Terms with $i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}$ correspond to subcubes to which we add mass to get the quantized masses in the k'-th coordinate, while terms with $i_{k'} = i_{k'}^{*(\mathbf{i}_{k'-1})}$ correspond to the subcube from which we take this extra mass. Correspondingly, we refer to terms with $i_{k'} \neq i_{k'}^{*(\mathbf{i}_{k'-1})}$ as "+ terms", while we designate terms with $i_{k'} = i_{k'}^{*(\mathbf{i}_{k'-1})}$ as "– terms". By construction, $(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}) \geq 0$ for + terms, while $(n_{\mathbf{i}_{k'}} - \tilde{n}_{\mathbf{i}_{k'}}) \geq 0$ for – terms. In evaluating the sum (26), we consider three different cases.

Case 1: All terms are + terms. In this case, the sum becomes

$$m_{\mathbf{i}_k} + \sum_{k'=1}^{k} \tilde{\Upsilon}(\mathbf{i}, k, k'+1)\big(\tilde{n}_{\mathbf{i}_{k'}} - n_{\mathbf{i}_{k'}}\big)\Upsilon(\mathbf{i}, k'-1, 1)$$

$$= m_{\mathbf{i}_k} + \sum_{k'=1}^{k} \tilde{\Upsilon}(\mathbf{i}, k, k')\Upsilon(\mathbf{i}, k'-1, 1) - \sum_{k'=1}^{k} \tilde{\Upsilon}(\mathbf{i}, k, k'+1)\Upsilon(\mathbf{i}, k', 1)$$

$$= m_{\mathbf{i}_k} - \underbrace{\Upsilon(\mathbf{i}, k, 1)}_{=m_{\mathbf{i}_k}} + \underbrace{\tilde{\Upsilon}(\mathbf{i}, k, 1)}_{=\tilde{m}_{\mathbf{i}_k}} + \sum_{k'=2}^{k} \tilde{\Upsilon}(\mathbf{i}, k, k')\Upsilon(\mathbf{i}, k'-1, 1) - \sum_{k'=1}^{k-1} \tilde{\Upsilon}(\mathbf{i}, k, k'+1)\Upsilon(\mathbf{i}, k', 1)$$

$$= \tilde{m}_{\mathbf{i}_k} + \sum_{k'=1}^{k-1} \tilde{\Upsilon}(\mathbf{i}, k, k'+1)\Upsilon(\mathbf{i}, k', 1) - \sum_{k'=1}^{k-1} \tilde{\Upsilon}(\mathbf{i}, k, k'+1)\Upsilon(\mathbf{i}, k', 1)$$

$$= \tilde{m}_{\mathbf{i}_k}.$$

Case 2: All terms are – terms. In this case, the sum is

$$m_{\mathbf{i}_k} - \sum_{k'=1}^{k} \Upsilon(\mathbf{i}, k, k'+1)\big(n_{\mathbf{i}_{k'}} - \tilde{n}_{\mathbf{i}_{k'}}\big)\tilde{\Upsilon}(\mathbf{i}, k'-1, 1)$$

$$= m_{\mathbf{i}_k} + \sum_{k'=1}^{k} \Upsilon(\mathbf{i}, k, k'+1)\tilde{\Upsilon}(\mathbf{i}, k', 1) - \sum_{k'=1}^{k} \Upsilon(\mathbf{i}, k, k')\tilde{\Upsilon}(\mathbf{i}, k'-1, 1)$$

$$= m_{\mathbf{i}_k} - \Upsilon(\mathbf{i}, k, 1) + \tilde{\Upsilon}(\mathbf{i}, k, 1) + \sum_{k'=1}^{k-1} \Upsilon(\mathbf{i}, k, k'+1)\tilde{\Upsilon}(\mathbf{i}, k', 1) - \sum_{k'=2}^{k} \Upsilon(\mathbf{i}, k, k')\tilde{\Upsilon}(\mathbf{i}, k'-1, 1)$$

$$= \tilde{m}_{\mathbf{i}_k}.$$

Case 3: There is at least one + term and one – term. Let the indices of the + terms be given by

$$\{k_1^+, \ldots, k_{\ell_1}^+\}$$

and those of the – terms by

$$\{k_1^-, \ldots, k_{\ell_2}^-\},$$

with both sets arranged in increasing order and $\ell_1 + \ell_2 = k$.

We first consider the sum of the – terms given by

$$\sum_{\ell=1}^{\ell_2} \Upsilon(\mathbf{i}, k, k_\ell^- + 1)\left(n_{\mathbf{i}_{k_\ell^-}} - \tilde{n}_{\mathbf{i}_{k_\ell^-}}\right)\Upsilon^\eta(\mathbf{i}, k_\ell^- - 1, 1) \tag{34}$$

and establish a cancelation property of successive terms in this sum, leaving only the border terms to be considered. Indeed, take ℓ such that $1 < \ell < \ell_2$, with the corresponding term given by

$$\Upsilon(\mathbf{i}, k, k_\ell^- + 1)\, n_{\mathbf{i}_{k_\ell^-}}\, \Upsilon^\eta(\mathbf{i}, k_\ell^- - 1, 1) \;-\; \Upsilon(\mathbf{i}, k, k_\ell^- + 1)\, \tilde{n}_{\mathbf{i}_{k_\ell^-}}\, \Upsilon^\eta(\mathbf{i}, k_\ell^- - 1, 1). \tag{35}$$

Next, note that the positive part of the term corresponding to the index $\ell + 1$ is given by

$$\Upsilon(\mathbf{i}, k, k_{\ell+1}^- + 1)\, n_{\mathbf{i}_{k_{\ell+1}^-}}\, \Upsilon^\eta(\mathbf{i}, k_{\ell+1}^- - 1, 1)$$
$$= \Upsilon(\mathbf{i}, k, k_{\ell+1}^- + 1)\, n_{\mathbf{i}_{k_{\ell+1}^-}}\, n_{\mathbf{i}_{k_{\ell+1}^- - 1}} \cdots n_{\mathbf{i}_{k_\ell^- + 1}}\, \tilde{n}_{\mathbf{i}_{k_\ell^-}}\, \Upsilon^\eta(\mathbf{i}, k_\ell^- - 1, 1), \tag{36}$$

since all indices that lie strictly between k_ℓ^- and $k_{\ell+1}^-$, if there are any, correspond to $+$ terms. Comparing (36) with (35) reveals that the positive part of the term corresponding to $\ell + 1$ cancels out the negative part of the term for ℓ. Similarly, the negative part of the term corresponding to $\ell - 1$, given by

$$-\Upsilon(\mathbf{i}, k, k_{\ell-1}^- + 1)\, \tilde{n}_{\mathbf{i}_{k_{\ell-1}^-}}\, \Upsilon^\eta(\mathbf{i}, k_{\ell-1}^- - 1, 1),$$

cancels out the positive part of the term for index ℓ, which is given by

$$\Upsilon(\mathbf{i}, k, k_\ell^- + 1)\, n_{\mathbf{i}_{k_\ell^-}}\, \Upsilon^\eta(\mathbf{i}, k_\ell^- - 1, 1)$$
$$= \Upsilon(\mathbf{i}, k, k_\ell^- + 1)\, n_{\mathbf{i}_{k_\ell^-}}\, n_{\mathbf{i}_{k_\ell^- - 1}} \cdots n_{\mathbf{i}_{k_{\ell-1}^- + 1}}\, \tilde{n}_{\mathbf{i}_{k_{\ell-1}^-}}\, \Upsilon^\eta(\mathbf{i}, k_{\ell-1}^- - 1, 1).$$

Hence, the only contributions remaining in the sum (34) over all the $-$ terms are the negative part of the term corresponding to the index ℓ_2 and the positive part of the term for the index 1, i.e.,

$$\sum_{\ell=1}^{\ell_2} \Upsilon(\mathbf{i}, k, k_\ell^- + 1)\left(n_{\mathbf{i}_{k_\ell^-}} - \tilde{n}_{\mathbf{i}_{k_\ell^-}} \right)\Upsilon^\eta(\mathbf{i}, k_\ell^- - 1, 1)$$
$$= \Upsilon(\mathbf{i}, k, k_1^- + 1)\, n_{\mathbf{i}_{k_1^-}}\, \Upsilon^\eta(\mathbf{i}, k_1^- - 1, 1) \;-\; \Upsilon(\mathbf{i}, k, k_{\ell_2}^- + 1)\, \tilde{n}_{\mathbf{i}_{k_{\ell_2}^-}}\, \Upsilon^\eta(\mathbf{i}, k_{\ell_2}^- - 1, 1)$$
$$= \Upsilon(\mathbf{i}, k, k_1^- + 1)\, n_{\mathbf{i}_{k_1^-}}\, \Upsilon(\mathbf{i}, k_1^- - 1, 1) \;-\; \Upsilon(\mathbf{i}, k, k_{\ell_2}^- + 1)\, \tilde{n}_{\mathbf{i}_{k_{\ell_2}^-}}\, \Upsilon^\eta(\mathbf{i}, k_{\ell_2}^- - 1, 1)$$
$$= \; m_{\mathbf{i}_k} - \Upsilon(\mathbf{i}, k, k_{\ell_2}^- + 1)\, \tilde{n}_{\mathbf{i}_{k_{\ell_2}^-}}\, \Upsilon^\eta(\mathbf{i}, k_{\ell_2}^- - 1, 1),$$

since indices smaller than k_1^- necessarily correspond to $+$ terms.

We proceed to the sum over the $+$ terms. A similar cancelation property between consecutive terms in the sum can be established so that we are again left with contributions from the first and the last term only. Indeed, take ℓ such that $1 < \ell < \ell_1$, with the corresponding term given by

$$\tilde{\Upsilon}(\mathbf{i}, k, k_\ell^+ + 1)\, \tilde{n}_{\mathbf{i}_{k_\ell^+}}\, \Upsilon^\eta(\mathbf{i}, k_\ell^+ - 1, 1) \;-\; \tilde{\Upsilon}(\mathbf{i}, k, k_\ell^+ + 1)\, n_{\mathbf{i}_{k_\ell^+}}\, \Upsilon^\eta(\mathbf{i}, k_\ell^+ - 1, 1).$$

The positive part of the term corresponding to the index $\ell + 1$ is given by

$$\tilde{\Upsilon}(\mathbf{i}, k, k_{\ell+1}^+ + 1)\, \tilde{n}_{\mathbf{i}_{k_{\ell+1}^+}}\, \Upsilon^\eta(\mathbf{i}, k_{\ell+1}^+ - 1, 1)$$
$$= \tilde{\Upsilon}(\mathbf{i}, k, k_{\ell+1}^+ + 1)\, \tilde{n}_{\mathbf{i}_{k_{\ell+1}^+}}\, \tilde{n}_{\mathbf{i}_{k_{\ell+1}^+ - 1}} \cdots \tilde{n}_{\mathbf{i}_{k_\ell^+ + 1}}\, n_{\mathbf{i}_{k_\ell^+}}\, \Upsilon^\eta(\mathbf{i}, k_\ell^+ - 1, 1),$$

since all indices that lie strictly between k_ℓ^+ and $k_{\ell+1}^+$, if there are any, correspond to $-$ terms. We can hence conclude, as above, that the positive part of the term corresponding to $\ell + 1$ cancels out the negative part of the term for ℓ. By the same argument, the positive part of the term for ℓ is cancelled out by the negative part of the term corresponding to $\ell - 1$. Overall, the only remaining contributions are the negative part of the term corresponding to ℓ_1 and the positive part of the term for the index 1, i.e.,

$$\sum_{\ell=1}^{\ell_1} \tilde{\Upsilon}(\mathbf{i}, k, k_\ell^+ + 1)\left(\tilde{n}_{\mathbf{i}_{k_\ell^+}} - n_{\mathbf{i}_{k_\ell^+}} \right)\Upsilon^\eta(\mathbf{i}, k_\ell^+ - 1, 1)$$
$$= \tilde{\Upsilon}(\mathbf{i}, k, 1) - \tilde{\Upsilon}(\mathbf{i}, k, k_{\ell_1}^+ + 1)\, n_{\mathbf{i}_{k_{\ell_1}^+}}\, \Upsilon^\eta(\mathbf{i}, k_{\ell_1}^+ - 1, 1)$$
$$= \tilde{m}_{\mathbf{i}_k} - \tilde{\Upsilon}(\mathbf{i}, k, k_{\ell_1}^+ + 1)\, n_{\mathbf{i}_{k_{\ell_1}^+}}\, \Upsilon^\eta(\mathbf{i}, k_{\ell_1}^+ - 1, 1).$$

33

Putting pieces together, (26) reduces to

$$
\begin{aligned}
& m_{\mathbf{i}_k} + \tilde{m}_{\mathbf{i}_k} - m_{\mathbf{i}_k} \\
& + \Upsilon(\mathbf{i}, k, k_{\ell_2}^- + 1)\, \tilde{n}_{\mathbf{i}_{k_{\ell_2}^-}} \Upsilon^\eta(\mathbf{i}, k_{\ell_2}^- - 1, 1) \\
& - \tilde{\Upsilon}(\mathbf{i}, k, k_{\ell_1}^+ + 1)\, n_{\mathbf{i}_{k_{\ell_1}^+}} \Upsilon^\eta(\mathbf{i}, k_{\ell_1}^+ - 1, 1).
\end{aligned}
\tag{37}
$$

There are two possibilities to consider now, either $k_{\ell_1}^+ = k$ or $k_{\ell_2}^- = k$. If $k_{\ell_1}^+ = k$, then (37) becomes

$$
\begin{aligned}
& \tilde{m}_{\mathbf{i}_k} \\
& + \Upsilon(\mathbf{i}, k_{\ell_1}^+, k_{\ell_2}^- + 1)\, \tilde{n}_{\mathbf{i}_{k_{\ell_2}^-}} \Upsilon^\eta(\mathbf{i}, k_{\ell_2}^- - 1, 1) \\
& - n_{\mathbf{i}_{k_{\ell_1}^+}} \cdots n_{\mathbf{i}_{k_{\ell_2}^- + 1}}\, \tilde{n}_{\mathbf{i}_{k_{\ell_2}^-}} \Upsilon^\eta(\mathbf{i}, k_{\ell_2}^- - 1, 1) \\
& = \tilde{m}_{\mathbf{i}_k},
\end{aligned}
$$

since, by definition, $k_{\ell_2}^-$ is the largest index corresponding to – terms. On the other hand, if $k_{\ell_2}^- = k$, then (37) reduces to

$$
\begin{aligned}
& \tilde{m}_{\mathbf{i}_k} \\
& + \tilde{n}_{\mathbf{i}_{k_{\ell_2}^-}} \cdots \tilde{n}_{\mathbf{i}_{k_{\ell_1}^+ + 1}}\, n_{\mathbf{i}_{k_{\ell_1}^+}} \Upsilon^\eta(\mathbf{i}, k_{\ell_1}^+ - 1, 1) \\
& - \tilde{\Upsilon}(\mathbf{i}, k_{\ell_2}^-, k_{\ell_1}^+ + 1)\, n_{\mathbf{i}_{k_{\ell_1}^+}} \Upsilon^\eta(\mathbf{i}, k_{\ell_1}^+ - 1, 1) \\
& = \tilde{m}_{\mathbf{i}_k},
\end{aligned}
$$

since, again by definition, $k_{\ell_1}^+$ is the largest index corresponding to + terms. This concludes the proof. $\quad\square$

References

[1] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: D. Precup, Y.W. Teh (eds.) Proceedings of the 34th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 70, pp. 214–223. PMLR (2017). URL http://proceedings.mlr.press/v70/arjovsky17a.html

[2] Bailey, B., Telgarsky, M.J.: Size-noise tradeoffs in generative networks. In: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (eds.) Advances in Neural Information Processing Systems 31, pp. 6489–6499. Curran Associates, Inc. (2018)

[3] Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory **39**(3), 930–945 (1993). DOI 10.1109/18.256500

[4] Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pp. 10–21. Association for Computational Linguistics, Berlin, Germany (2016). DOI 10.18653/v1/K16-1002. URL https://www.aclweb.org/anthology/K16-1002

[5] Box, G.E.P., Muller, M.E.: A note on the generation of random normal deviates. Ann. Math. Statist. **29**(2), 610–611 (1958). DOI 10.1214/aoms/1177706645. URL https://doi.org/10.1214/aoms/1177706645

[6] Bölcskei, H., Grohs, P., Kutyniok, G., Petersen, P.: Optimal approximation with sparsely connected deep neural networks. SIAM Journal on Mathematics of Data Science **1**(1), 8–45 (2019). URL https://www.mins.ee.ethz.ch/pubs/p/deep-approx-18

[7] Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems **2**(4), 303–314 (1989). DOI 10.1007/BF02551274. URL https://doi.org/10.1007/BF02551274

[8] Daubechies, I., DeVore, R., Foucart, S., Hanin, B., Petrova, G.: Nonlinear approximation and (deep) ReLU networks. Constructive Approximation (2021). DOI 10.1007/s00365-021-09548-z. URL https://doi.org/10.1007/s00365-021-09548-z

[9] Elbrächter, D., Perekrestenko, D., Grohs, P., Bölcskei, H.: Deep neural network approximation theory. IEEE Transactions on Information Theory **67**(5), pp. 2581–2623 (2021). URL https://www.mins.ee.ethz.ch/pubs/p/deep-it-2019

[10] Eldan, R., Shamir, O.: The power of depth for feedforward neural networks. In: V. Feldman, A. Rakhlin, O. Shamir (eds.) 29th Annual Conference on Learning Theory, *Proceedings of Machine Learning Research*, vol. 49, pp. 907–940. PMLR, Columbia University, New York, New York, USA (2016). URL http://proceedings.mlr.press/v49/eldan16.html

[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014). URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[12] Graf, S., Luschgy, H.: Foundations of Quantization for Probability Distributions. Springer-Verlag, Berlin, Heidelberg (2000)

[13] Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks **2**(5), 359 – 366 (1989). DOI https://doi.org/10.1016/0893-6080(89)90020-8. URL http://www.sciencedirect.com/science/article/pii/0893608089900208

[14] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401–4410 (2019)

[15] Kingma, D., Welling, M.: Auto-encoding variational Bayes. Proceedings of the International Conference on Learning Representations (ICLR) (2014)

[16] Lee, H., Ge, R., Risteski, A., Ma, T., Arora, S.: On the ability of neural nets to express distributions. Proceedings of Machine Learning Research **65**, 1–26 (2017)

[17] Lu, Y., Lu, J.: A universal approximation theorem of deep neural networks for expressing probability distributions. In: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 3094–3105. Curran Associates, Inc. (2020). URL https://proceedings.neurips.cc/paper/2020/file/2000f6325dfc4fc3201fc45ed01c7a5d-Paper.pdf

[18] McCann, R.J., Pass, B.: Optimal transportation between unequal dimensions. Archive for Rational Mechanics and Analysis (2020). DOI 10.1007/s00205-020-01569-5

[19] Perekrestenko, D., Müller, S., Bölcskei, H.: Constructive universal high-dimensional distribution generation through deep ReLU networks. In: Proc. of the 37th International Conference on Machine Learning (ICML) (2020). URL https://www.mins.ee.ethz.ch/pubs/p/icml2020

[20] Peyré, G., Cuturi, M.: Computational optimal transport. Foundations and Trends in Machine Learning **11**(5-6), 355–607 (2019). DOI 10.1561/2200000073. URL http://dx.doi.org/10.1561/2200000073

[21] Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. Proceedings of the International Conference on Learning Representations (ICLR) (2016)

[22] Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: F. Bach, D. Blei (eds.) Proceedings of the 32nd International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 37, pp. 1530–1538. PMLR, Lille, France (2015)

[23] Telgarsky, M.: Benefits of depth in neural networks. JMLR: Workshop and Conference Proceedings **49**, 1–23 (2016)

[24] Tolstikhin, I., Bousquet, O., Gelly, S., Schölkopf, B.: Wasserstein auto-encoders. In: 6th International Conference on Learning Representations (ICLR) (2018). URL https://openreview.net/forum?id=HkL7n1-0b

[25] Villani, C.: Optimal transport: Old and new, vol. 338. Springer Science & Business Media (2008)

[26] Xu, J., Ren, X., Lin, J., Sun, X.: Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3940–3949. Association for Computational Linguistics, Brussels, Belgium (2018). DOI 10.18653/v1/D18-1428. URL https://www.aclweb.org/anthology/D18-1428