# Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity

**Chulhee Yun, Suvrit Sra, Ali Jadbabaie**

Laboratory for Information and Decision Systems, MIT

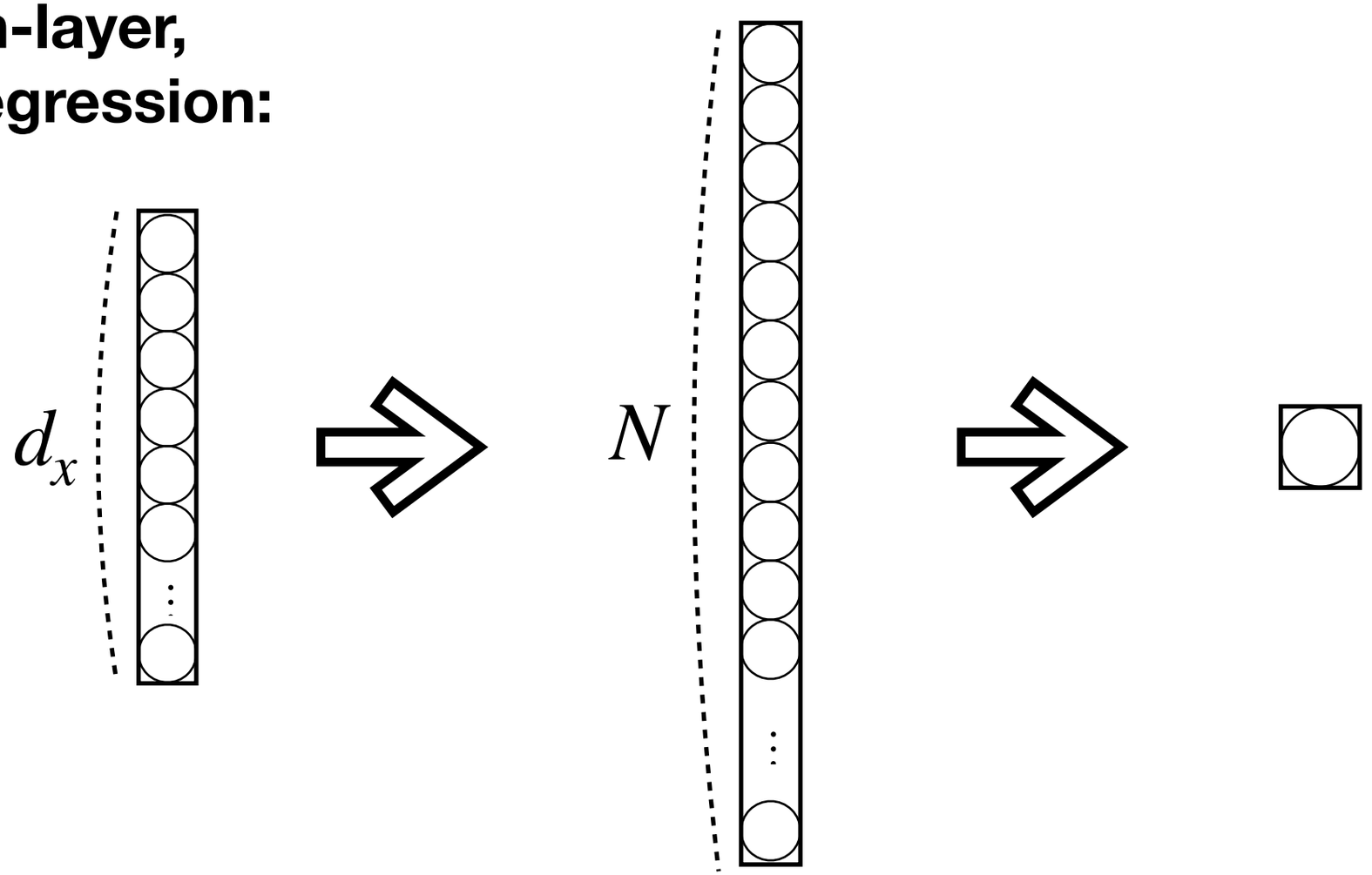
Given a ReLU fully-connected network,

**how many hidden nodes are required to memorize arbitrary $N$ data points?**

Given a ReLU fully-connected network,

**how many hidden nodes are required to memorize arbitrary $N$ data points?**

**1-hidden-layer, scalar regression:**



We prove that for 2-hidden-layer networks,

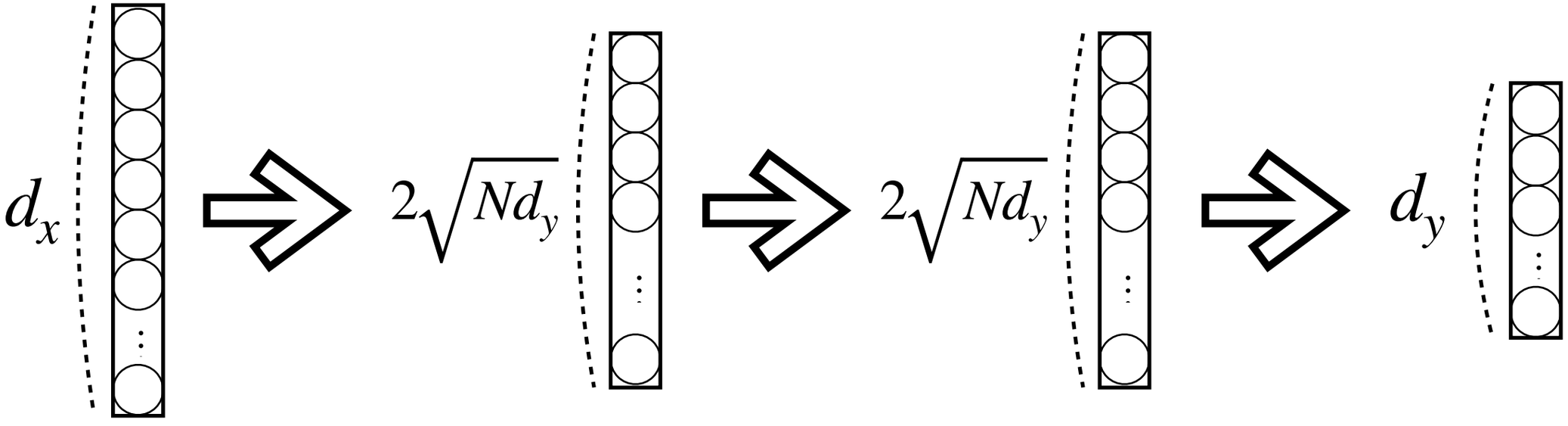
$\Theta(\sqrt{Nd_y})$ neurons are **sufficient**.

If $d_y = 1$, $\Theta(\sqrt{N})$ neurons are also **necessary**.

We prove that for 2-hidden-layer networks,

$\Theta(\sqrt{Nd_y})$ neurons are **sufficient**.

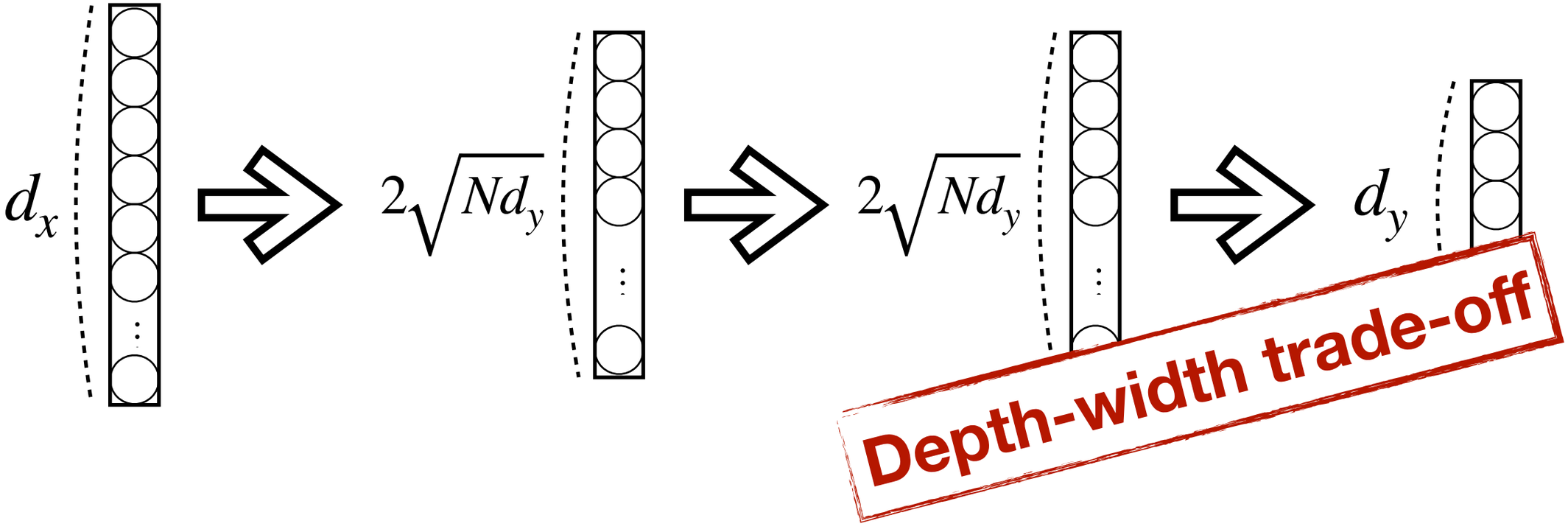
If $d_y = 1$, $\Theta(\sqrt{N})$ neurons are also **necessary**.



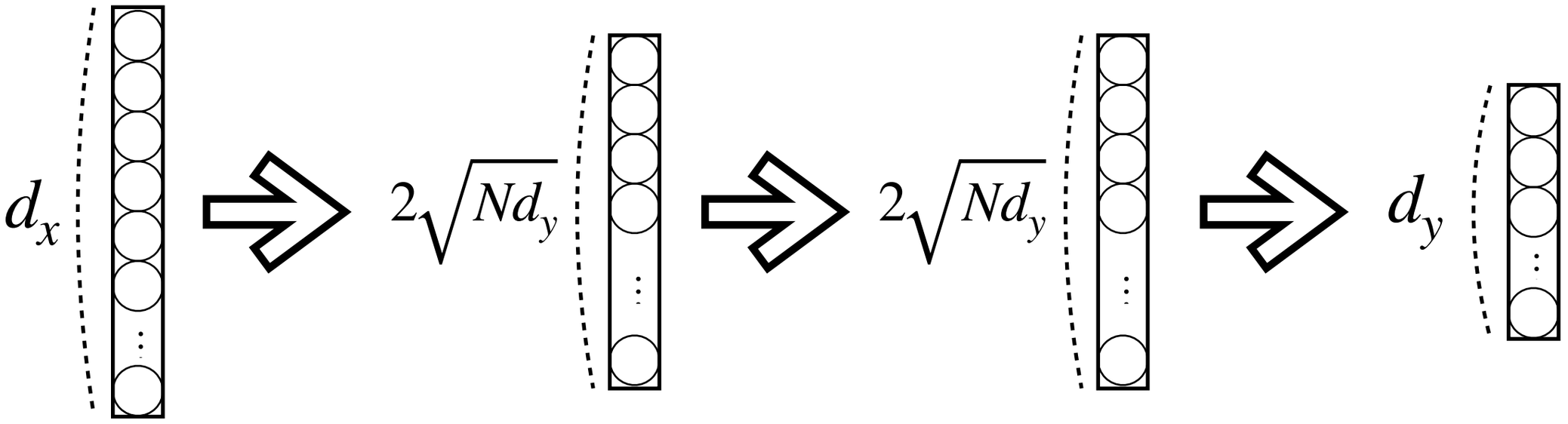
We prove that for 2-hidden-layer networks,

$\Theta(\sqrt{Nd_y})$ neurons are **sufficient**.

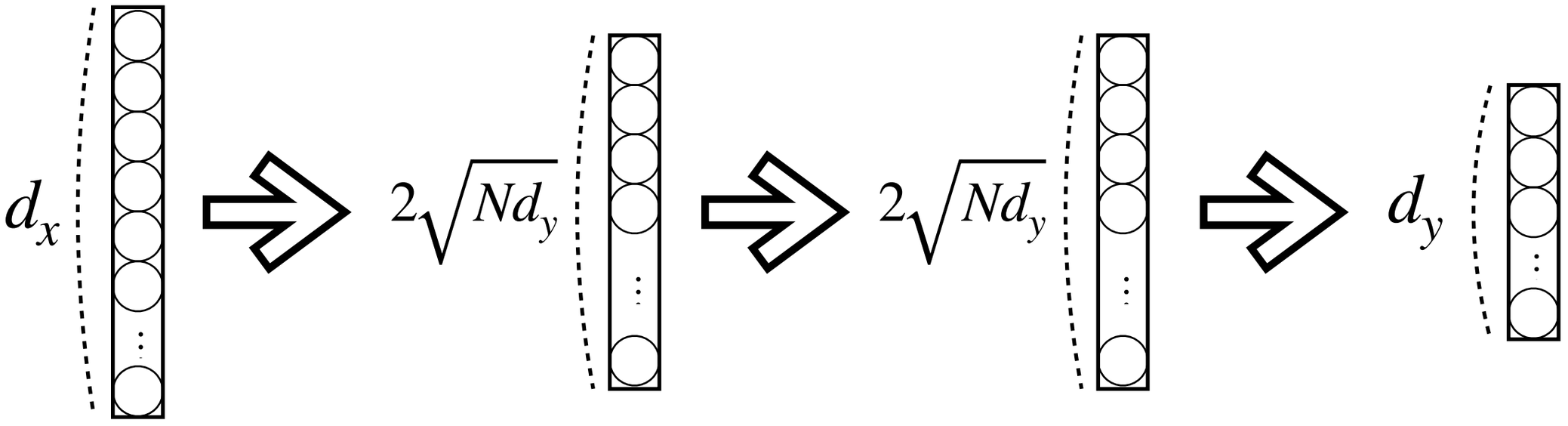
If $d_y = 1$, $\Theta(\sqrt{N})$ neurons are also **necessary**.



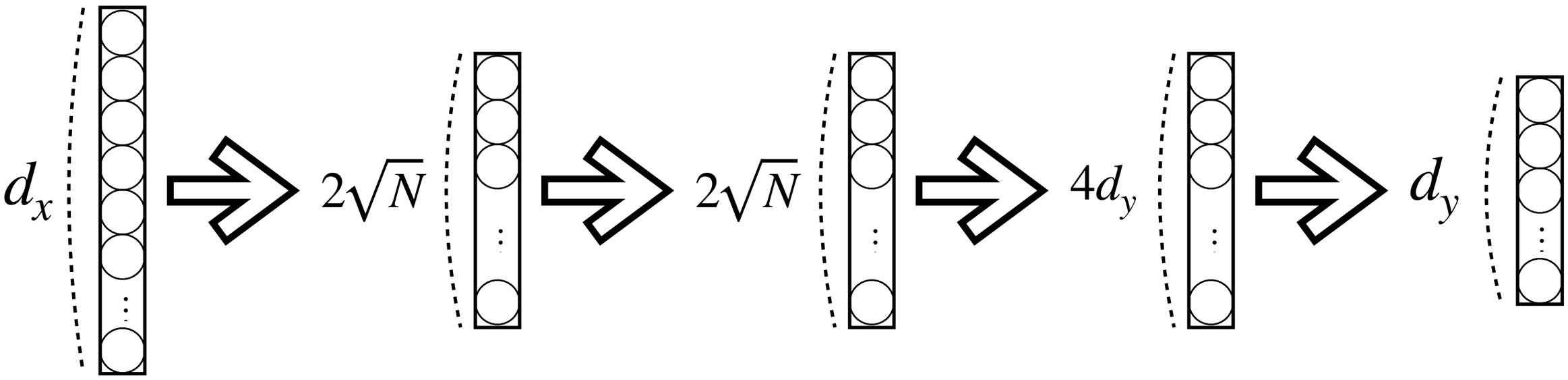
**Regression:**



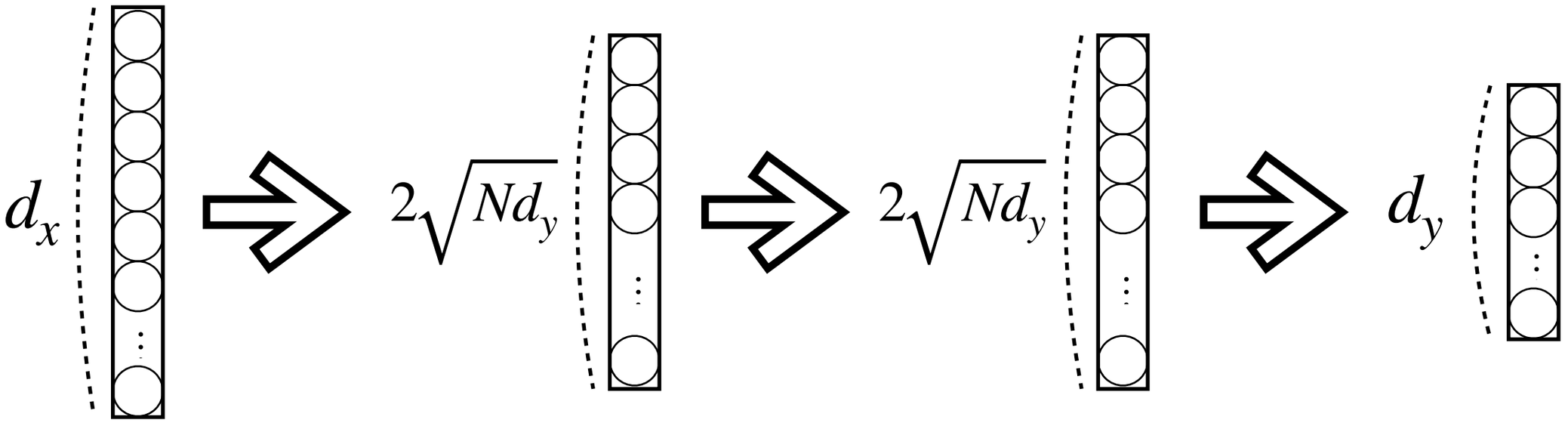
**Regression:**



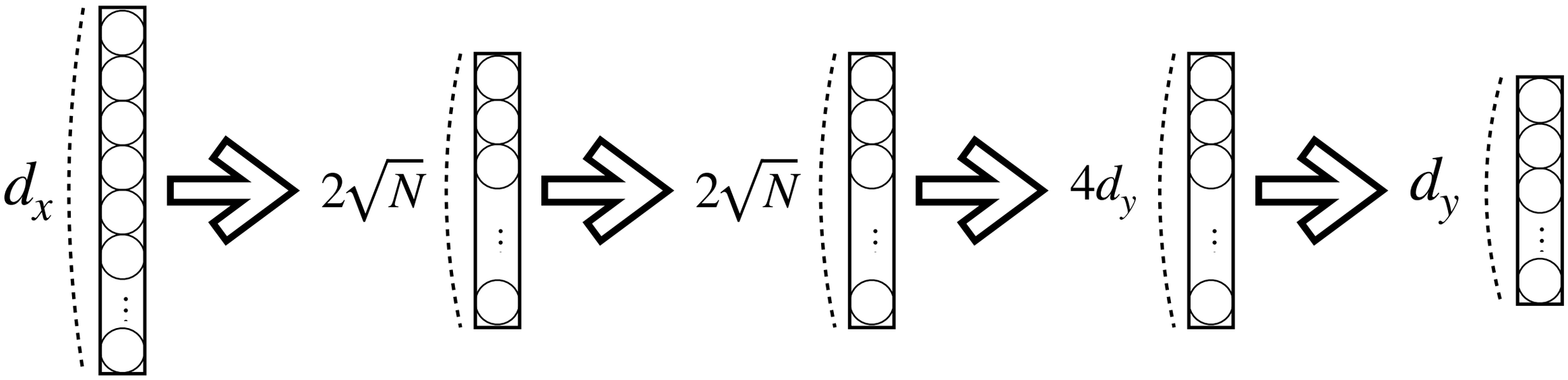

**Classification:**



**Regression:**

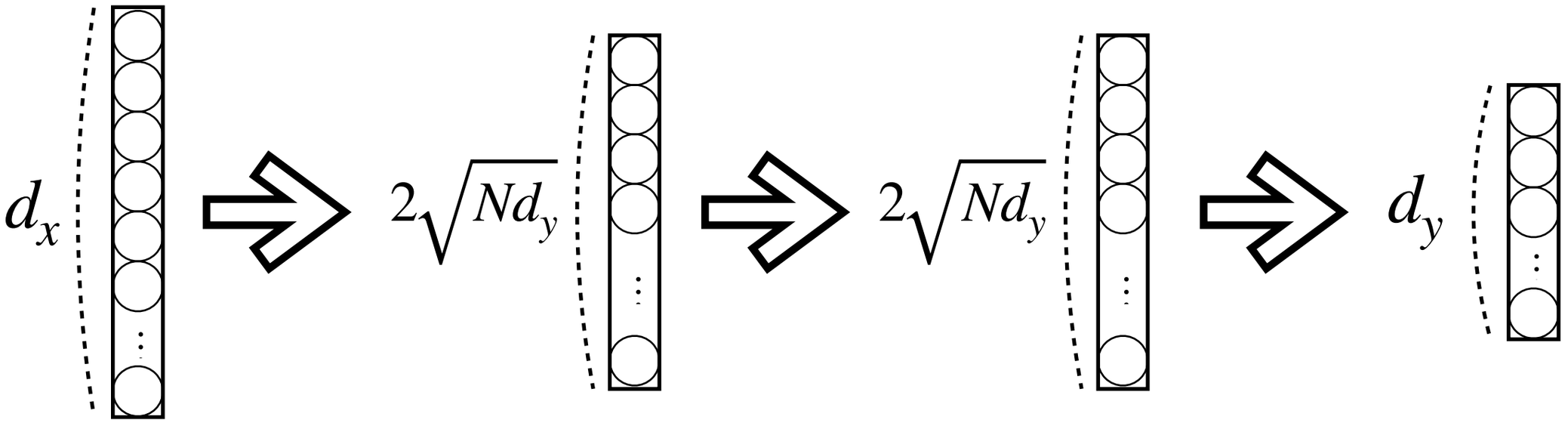



**Classification:**

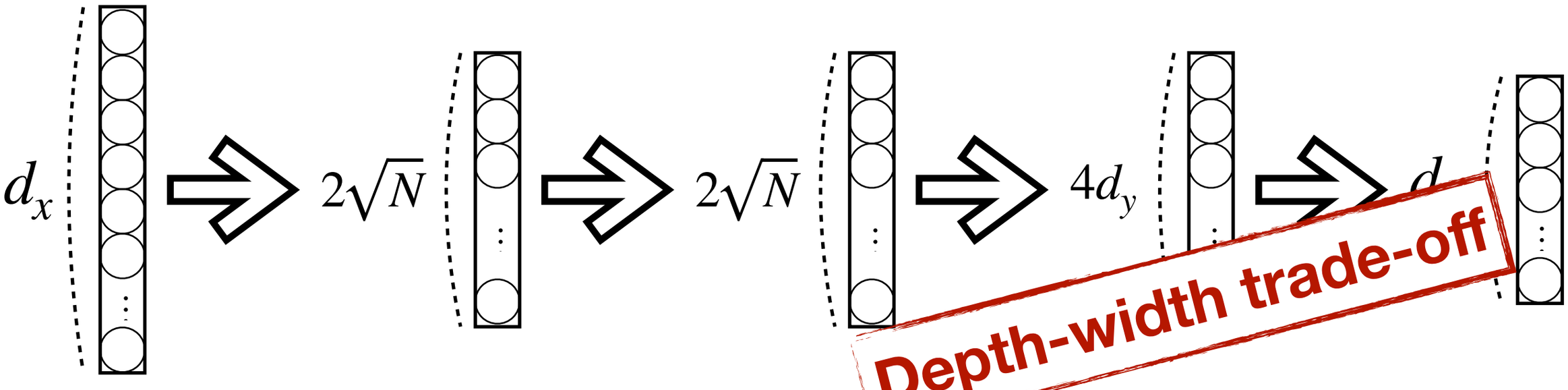



**ImageNet ($N = 1\text{M}$, $d_y = 1\text{k}$) memorized with 2k-2k-4k**

**Regression:**

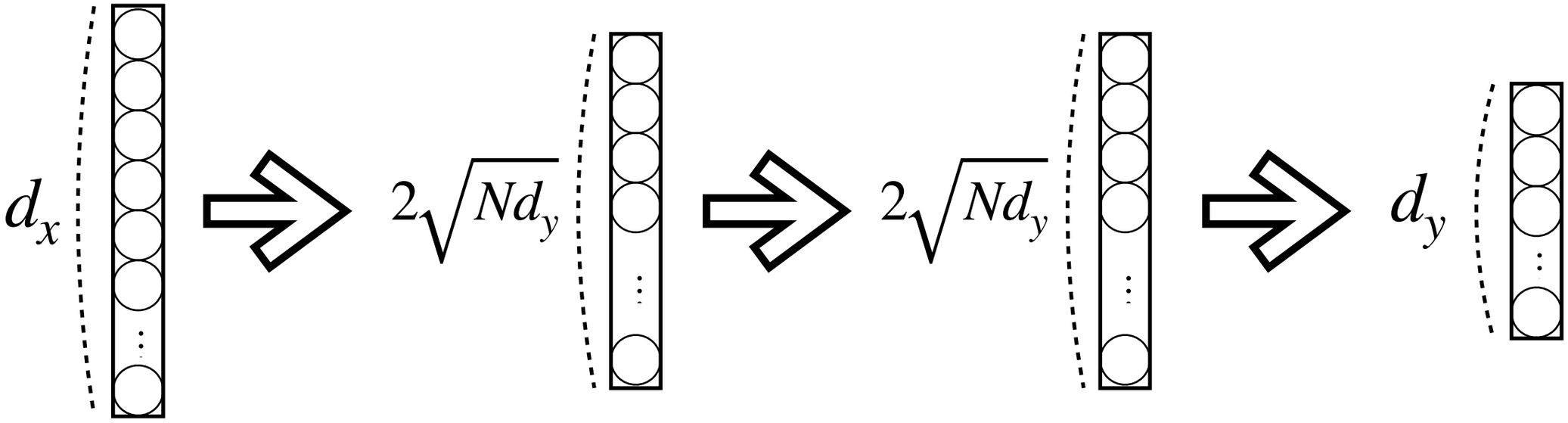



**Classification:**

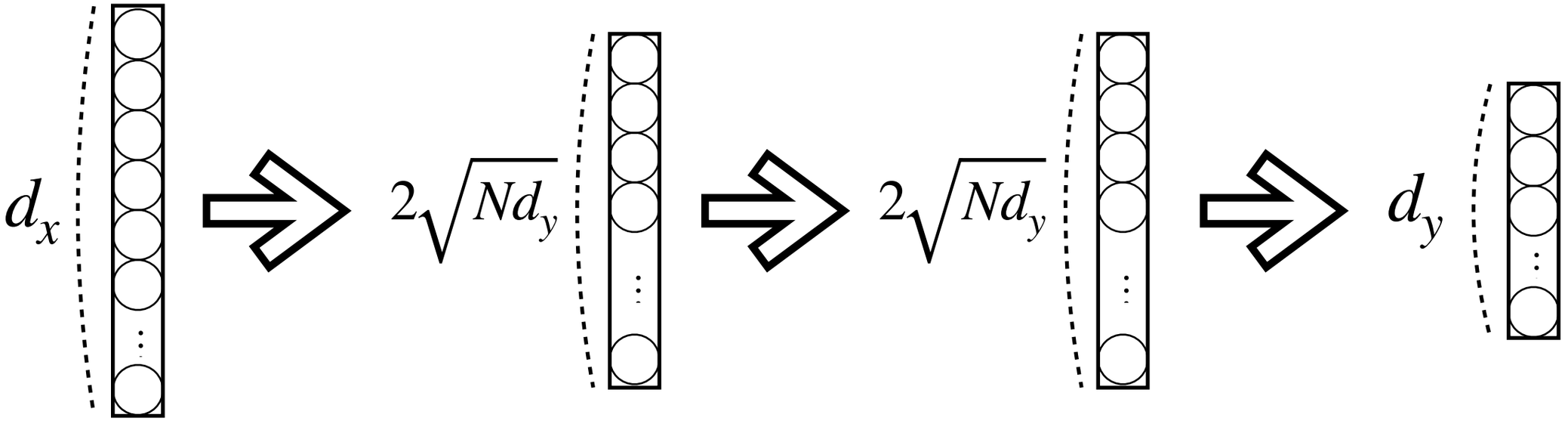



**ImageNet ($N = 1\text{M}$, $d_y = 1\text{k}$) memorized with 2k-2k-4k**

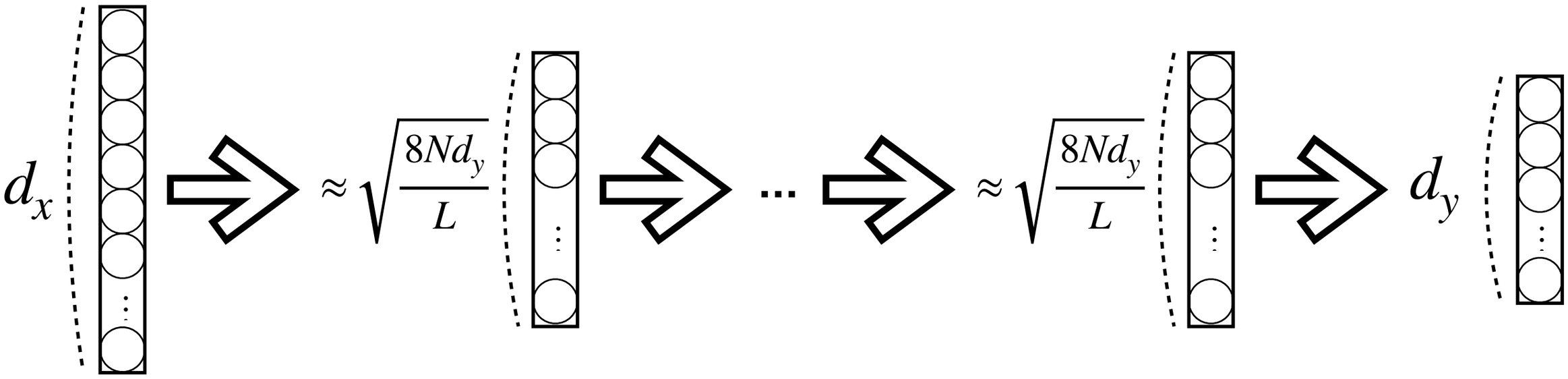
## 2 hidden layers:



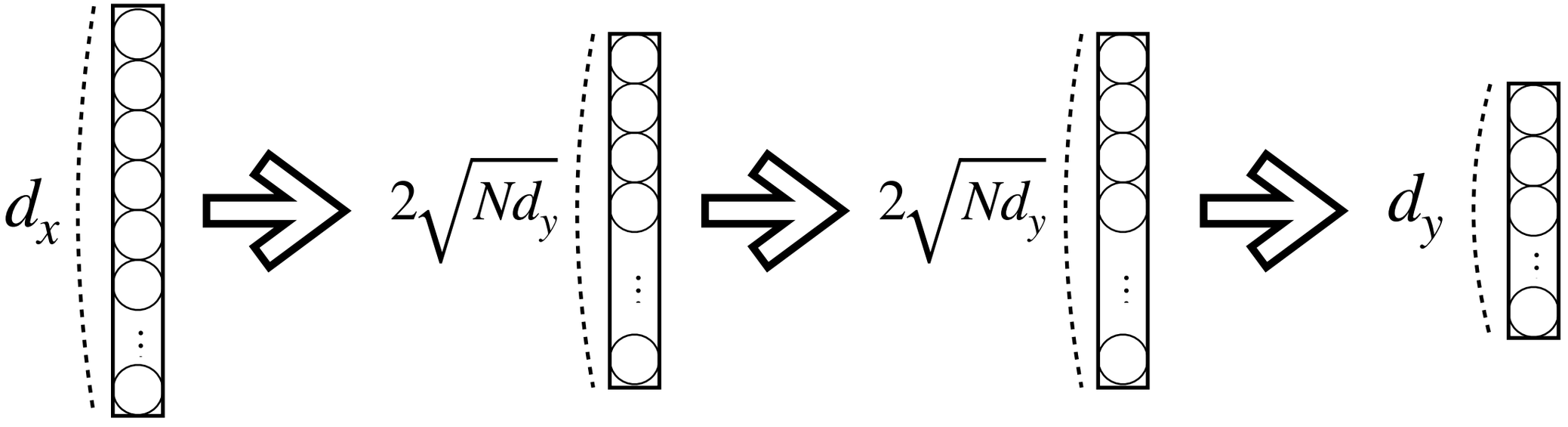
**2 hidden layers:**



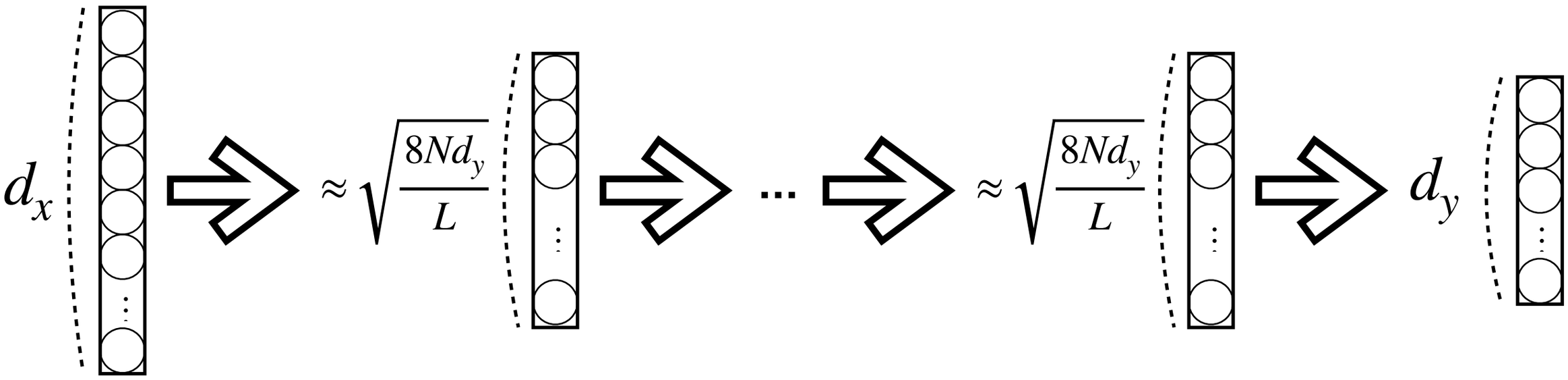

**$L$ hidden layers:**



**2 hidden layers:**




**$L$ hidden layers:**




**A Network with $W$ params can memorize if $W = \Omega(N)$**

Given a network, we define **memorization capacity** $C$ as

$C = \max\{N \mid$ the network can memorize arbitrary $N$ data points with $d_y = 1\}$

Given a network, we define **memorization capacity** $C$ as

$C = \max\{N \mid \text{the network can memorize arbitrary } N \text{ data points with } d_y = 1\}$

$\Theta(\sqrt{N})$ neurons necessary and sufficient for 2-hidden-layer

$$\Longrightarrow \quad C = \Theta(W)$$

Given a network, we define **memorization capacity** $C$ as

$C = \max\{N \mid$ the network can memorize arbitrary $N$ data points with $d_y = 1\}$

$\Theta(\sqrt{N})$ neurons necessary and sufficient for 2-hidden-layer

$$\Longrightarrow \quad C = \Theta(W)$$

Tight

Given a network, we define **memorization capacity** $C$ as

$C = \max\{N \mid$ the network can memorize arbitrary $N$ data points with $d_y = 1\}$

$\Theta(\sqrt{N})$ neurons necessary and sufficient for 2-hidden-layer

$$\implies C = \Theta(W)$$

Tight

$W = \Omega(N)$ sufficient for $L$-hidden-layer

$$\implies C = \Omega(W)$$

$C \leq \text{VCdim} = O(WL \log W)$

Given a network, we define **memorization capacity** $C$ as

$C = \max\{N \mid$ the network can memorize arbitrary $N$ data points with $d_y = 1\}$

$\Theta(\sqrt{N})$ neurons necessary and sufficient for 2-hidden-layer

$\implies C = \Theta(W)$

Tight

$W = \Omega(N)$ sufficient for $L$-hidden-layer

$\implies C = \Omega(W)$

Nearly Tight

$C \leq \text{VCdim} = O(WL \log W)$

## Other results

- Tighter sufficient condition for memorizing in residual network

- SGD trajectory analysis near memorizing global minimum

# Poster #233

# Wed Dec 11th 5PM-7PM

# @ East Exhibition Hall B + C