VC-dimension of neural networks

Abbas Mehrabian

McGill University
IVADO Postdoctoral Fellow

29 January 2019

Binary classification tasks

- ✓ Emails: spam/not spam
- ✓ Mammograms: cancerous/non-cancerous

Input to learning algorithm: a set of examples labelled by an expert.

Question. How many labelled examples are needed for learning a model? (sample complexity)

Binary classification

Domain $\mathcal{X}$
Distribution D over $\mathcal{X}$
true classifier $t : \mathcal{X} \to \{-1, +1\}$

Goal of learning: output some $h : \mathcal{X} \to \{-1, +1\}$ with small error

$$\text{error}(h) := \mathbf{P}_{X \sim D}\{h(X) \neq t(X)\}$$

Input: $X_1, \ldots, X_m \sim D$ and $t(X_1), \ldots, t(X_m)$

Empirical Risk Minimization (ERM)

$\text{error}(h) := \mathbf{P}_{X\sim D}\{h(X) \neq t(X)\}$ and $X_1, \ldots, X_m \sim D$.

Choose some class H of functions $\mathcal{X} \to \{-1, +1\}$, and output

$$h^* = \arg\min_{h \in H} \underbrace{\sum_{i=1}^{m} \frac{1}{m} \mathbb{1}\left[h(X_i) \neq t(X_i)\right]}_{\text{training error of } h}$$

$\sum_{i=1}^{m} \frac{1}{m} \mathbb{1}\left[h(X_i) \neq t(X_i)\right]$ is the empirical estimate for $\text{error}(h)$.

Empirical Risk Minimization (ERM)

$\text{error}(h) := \mathbf{P}_{X \sim D}\{h(X) \neq t(X)\}$ and $X_1, \ldots, X_m \sim D$.

Choose some class H of functions $\mathcal{X} \to \{-1, +1\}$, and output

$$h^* = \arg\min_{h \in H} \underbrace{\sum_{i=1}^{m} \frac{1}{m} \mathbb{1}\left[h(X_i) \neq t(X_i)\right]}_{\text{training error of } h}$$

$\sum_{i=1}^{m} \frac{1}{m} \mathbb{1}\left[h(X_i) \neq t(X_i)\right]$ is the empirical estimate for $\text{error}(h)$.

$\text{error}(h^*)$ = training error + estimation error (generalization error)
(Bias-complexity trade-off)

Bounding the estimation error

The case of finite H

Fix $h \in H$. Recall $\text{error}(h) \coloneqq \mathbf{P}_{X \sim D}\{h(X) \neq t(X)\}$ and $X_i \sim D$.
Then $\mathbb{1}[h(X_i) \neq t(X_i)]$ is Bernoulli with parameter $\text{error}(h)$.
Hoeffding's inequality:

$$\mathbb{P}\left[\underbrace{\sum_{i=1}^{m} \frac{1}{m} \mathbb{1}[h(X_i) \neq t(X_i)] - \text{error}(h)}_{\text{estimation error of h}} > t\right] < \exp(-2mt^2),$$

Bounding the estimation error

The case of finite H

Fix $h \in H$. Recall $\text{error}(h) := \mathbf{P}_{X \sim D}\{h(X) \neq t(X)\}$ and $X_i \sim D$.
Then $\mathbb{1}[h(X_i) \neq t(X_i)]$ is Bernoulli with parameter $\text{error}(h)$.
Hoeffding's inequality:

$$\mathbb{P}\left[\underbrace{\sum_{i=1}^{m} \frac{1}{m}\mathbb{1}[h(X_i) \neq t(X_i)] - \text{error}(h)}_{\text{estimation error of h}} > t\right] < \exp(-2mt^2),$$

so

$$\mathbb{P}\left[\sup_{h \in H}\{\text{estimation error of } h\} > t\right] < |H| \exp(-2mt^2),$$

so

$$\mathbb{E}\left[\sup_{h \in H} \text{estimation error of } h\right] \leq 3\sqrt{\ln |H| / m}$$

VC-dimension of H

Say $x_1, \ldots, x_m \in \mathcal{X}$ is shattered if

$$|\{(h(x_1), h(x_2), \ldots, h(x_m)) : h \in H\}| = 2^m$$

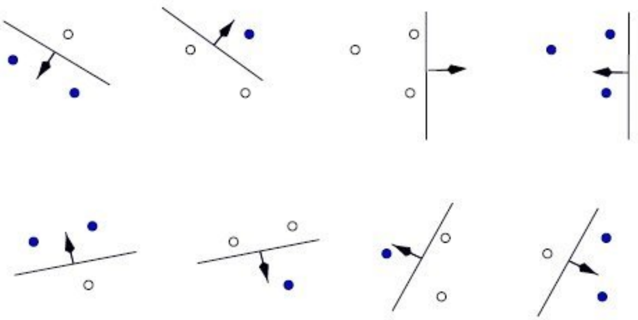
VC-dim$(H) :=$ size of the largest shattered set

Example: linear classifiers

Let $\mathcal{X} = \mathbb{R}^d$. A linear classifier is parametrized by $w_1, \ldots, w_{d+1}$:

$$h(x_1, \ldots, x_d) = \text{sgn}(w_1 x_1 + \cdots + w_d x_d + w_{d+1})$$

The VC-dimension of linear classifiers in $\mathbb{R}^2$ is 3

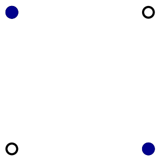


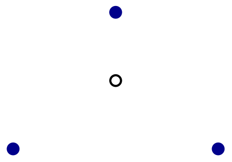
Example: linear classifiers

Let $\mathcal{X} = \mathbb{R}^d$. A linear classifier is parametrized by $w_1, \ldots, w_{d+1}$:

$$h(x_1, \ldots, x_d) = \operatorname{sgn}(w_1 x_1 + \cdots + w_d x_d + w_{d+1})$$

The VC-dimension of linear classifiers in $\mathbb{R}^2$ is 3





Example: linear classifiers

Let $\mathcal{X} = \mathbb{R}^d$. A linear classifier is parametrized by $w_1, \ldots, w_{d+1}$:

$$h(x_1, \ldots, x_d) = \mathrm{sgn}(w_1 x_1 + \cdots + w_d x_d + w_{d+1})$$

The VC-dimension of linear classifiers in $\mathbb{R}^2$ is 3

The VC-dimension of linear classifiers in $\mathbb{R}^d$ is $d + 1$

The Vapnik-Chervonenkis inequality

Recall: given a hypothesis class H, for any $h \in H$,
$\text{error}(h) = \text{training error} + \text{estimation error}$

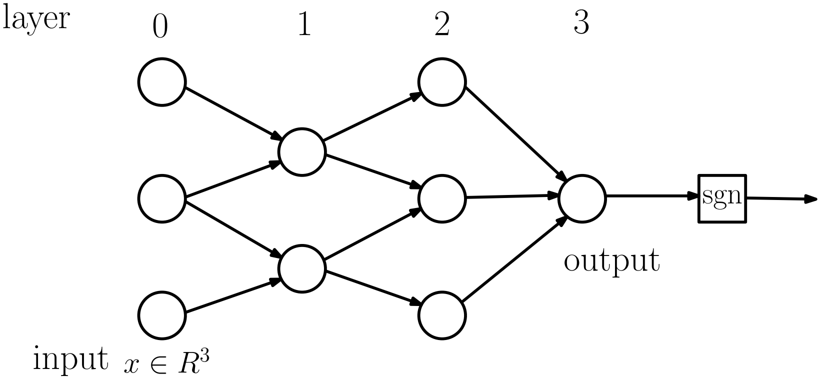
Theorem (Vapnik and Chervonenkis 1971, Dudley 1978)

Given a training set of size m,

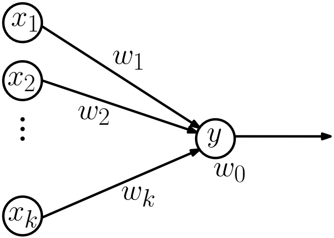
$$\mathbb{E}\left[\sup_{h \in H} \text{ estimation error of } h\right] \leq C\sqrt{\frac{\text{VC-dim}(H)}{m}}$$

This inequality is tight up to the value of C

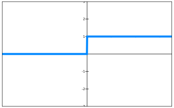
Neural networks



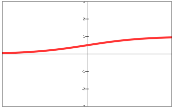
Neural networks



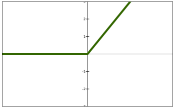
$$y = \sigma(w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_k x_k)$$



threshold
$\mathrm{sgn}(x)$

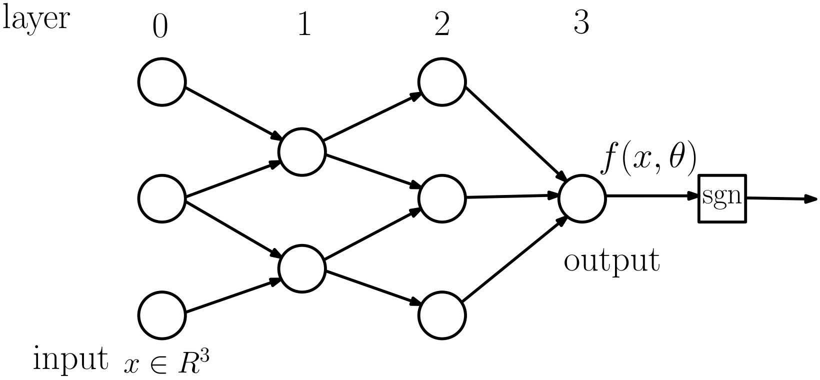


sigmoid
$\tanh(x)$

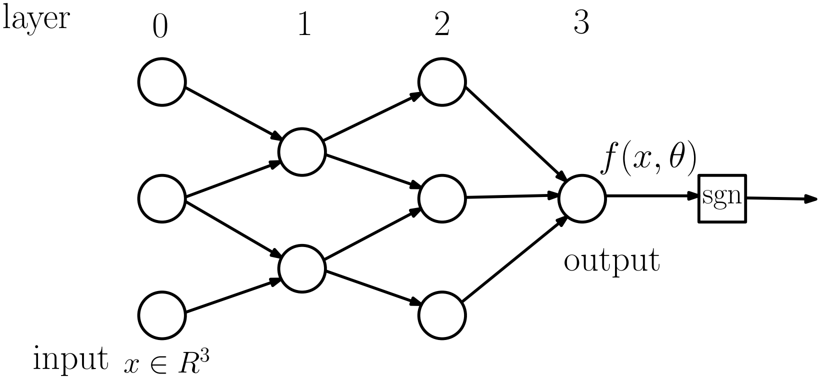


ReLU
$\max\{0, x\}$

Neural networks



Neural networks



Given a network architecture and activation functions, we obtain a class H of functions

$$H = \{\operatorname{sgn} f(x,\theta) : \theta \in \mathbb{R}^p\}$$

Main result

Theorem (Bartlett, Harvey, Liaw, M'17)

Let σ be a piecewise linear function, p be the number of parameters, ℓ the number of layers.
Then VC-dim $\leq O(p\ell \log p)$.
There exist networks with VC-dim $\geq \Omega(p\ell \log(p/\ell))$.

Main result

Theorem (Bartlett, Harvey, Liaw, M'17)

Let σ be a piecewise linear function, p be the number of parameters, ℓ the number of layers.
Then VC-dim $\leq O(p\ell \log p)$.
There exist networks with VC-dim $\geq \Omega(p\ell \log(p/\ell))$.

previous upper bounds: $O(p^2)$ [Goldberg and Jerrum'95]
$O(p\ell^2 + p\ell \log p)$ [Bartlett, Maiorov, Meir'98]

previous lower bounds: $\Omega(p \log p)$ [Maass'94]
$\Omega(p\ell)$ [Bartlett, Maiorov, Meir'98]

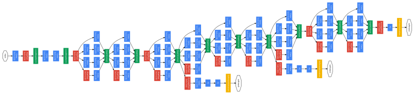
Main result

Theorem (Bartlett, Harvey, Liaw, M'17)

Let σ be a piecewise linear function, p be the number of parameters, ℓ the number of layers.
Then VC-dim $\leq O(p\ell \log p)$.
There exist networks with VC-dim $\geq \Omega(p\ell \log(p/\ell))$.

previous upper bounds: $O(p^2)$ [Goldberg and Jerrum'95]
$O(p\ell^2 + p\ell \log p)$ [Bartlett, Maiorov, Meir'98]

previous lower bounds: $\Omega(p \log p)$ [Maass'94]
$\Omega(p\ell)$ [Bartlett, Maiorov, Meir'98]



GoogleNet 2014: $\ell = 22, p = 4$ million

Upper bound proof

Assume all biases are 0, all layers have k nodes, $\sigma(x) = \max\{0, x\}$. So $p = (\ell - 1)k^2 + k$. Let $x_1, \dots, x_m$ be shattered.

$$\begin{aligned} 2^m &= |\{(\operatorname{sgn} f(x_1, \theta), \dots, \operatorname{sgn} f(x_m, \theta)) : \theta \in \mathbb{R}^p\}| \\ &= |\{(\operatorname{sgn} g_1(\theta), \dots, \operatorname{sgn} g_m(\theta)) : \theta \in \mathbb{R}^p\}| = \Pi \end{aligned}$$

Lemma (Warren'68)

If $q_1, \dots, q_m$ are polynomials of degree d in n variables,

$$|\{(\operatorname{sgn} q_1(y), \dots, \operatorname{sgn} q_m(y)) : y \in \mathbb{R}^n\}| \leq (6md/n)^n$$

Upper bound proof

Assume all biases are 0, all layers have k nodes, $\sigma(x) = \max\{0, x\}$. So $p = (\ell - 1)k^2 + k$. Let $x_1, \ldots, x_m$ be shattered.

$$\begin{aligned} 2^m &= |\{(\operatorname{sgn} f(x_1, \theta), \ldots, \operatorname{sgn} f(x_m, \theta)) : \theta \in \mathbb{R}^p\}| \\ &= |\{(\operatorname{sgn} g_1(\theta), \ldots, \operatorname{sgn} g_m(\theta)) : \theta \in \mathbb{R}^p\}| = \Pi \end{aligned}$$

Lemma (Warren'68)

If $q_1, \ldots, q_m$ are polynomials of degree d in n variables,

$$|\{(\operatorname{sgn} q_1(y), \ldots, \operatorname{sgn} q_m(y)) : y \in \mathbb{R}^n\}| \leq (6md/n)^n$$

Imagine that each $g_j(\theta)$ was a polynomial of degree ℓ in the p weights (variables). Then Warren's lemma would give

$$2^m = \Pi \leq (6m\ell/p)^p \leq (6m)^p = 2^{p \log_2(6m)},$$

so $m \leq p \log_2(6m) \Rightarrow m \leq O(p \log p)$

Upper bound proof

Assume all biases are 0, all layers have k nodes, $\sigma(x) = \max\{0, x\}$. So $p = (\ell - 1)k^2 + k$. Let $x_1, \ldots, x_m$ be shattered.

$$\begin{aligned} 2^m &= |\{(\operatorname{sgn} f(x_1, \theta), \ldots, \operatorname{sgn} f(x_m, \theta)) : \theta \in \mathbb{R}^p\}| \\ &= |\{(\operatorname{sgn} g_1(\theta), \ldots, \operatorname{sgn} g_m(\theta)) : \theta \in \mathbb{R}^p\}| = \Pi \end{aligned}$$

Lemma (Warren'68)

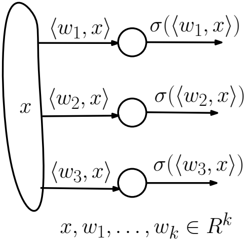
If $q_1, \ldots, q_m$ are polynomials of degree d in n variables,

$$|\{(\operatorname{sgn} q_1(y), \ldots, \operatorname{sgn} q_m(y)) : y \in \mathbb{R}^n\}| \leq (6md/n)^n$$

Instead, we build an iterative sequence of refined partitions $\mathcal{S}_1, \ldots, \mathcal{S}_{\ell-1}$ of $\mathbb{R}^p$, so that each $g_j(\theta)$ is a piecewise polynomial of degree ℓ with pieces given by $\mathcal{S}_{\ell-1}$, so

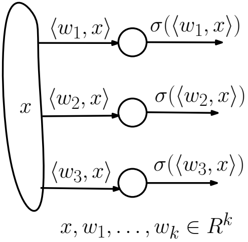
$$2^m = \Pi \leq (6m\ell/p)^p \cdot |\mathcal{S}_{\ell-1}|$$

Iterative construction of the partitions



1. $\{\langle w_i, x_j\rangle : i = 1, \ldots, k, j = 1, \ldots, m\}$ is a collection of km linear polynomials in k^2 variables of $w_1, \ldots, w_k$. By Warren, number of attained sign vectors is $\leq (6km/k^2)^{k^2} = (6m/k)^{k^2}$.

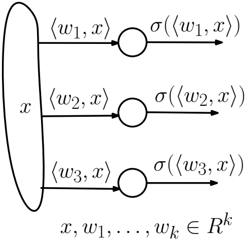
Iterative construction of the partitions



1. $\{\langle w_i, x_j \rangle : i = 1, \ldots, k, j = 1, \ldots, m\}$ is a collection of km linear polynomials in k^2 variables of $w_1, \ldots, w_k$. By Warren, number of attained sign vectors is $\leq (6km/k^2)^{k^2} = (6m/k)^{k^2}$.

2. Let $\mathcal{S}_1$ =partition specified by this sign pattern. $|\mathcal{S}_1| \leq (6m/k)^{k^2}$.

Iterative construction of the partitions



1. $\{\langle w_i, x_j\rangle : i = 1, \dots, k, j = 1, \dots, m\}$ is a collection of km linear polynomials in k^2 variables of $w_1, \dots, w_k$. By Warren, number of attained sign vectors is $\leq (6km/k^2)^{k^2} = (6m/k)^{k^2}$.

2. Let $\mathcal{S}_1$ =partition specified by this sign pattern. $|\mathcal{S}_1| \leq (6m/k)^{k^2}$.

3. Within each part of $\mathcal{S}_1$, the sign vector $(\mathrm{sgn}\langle w_i, x_j\rangle)_{i,j}$ is fixed, so $(\sigma\langle w_i, x_j\rangle)_{i,j}$ are linear polynomials in k^2 variables, so the inputs to the second layer are quadratic polynomials in $2k^2$ parameters.

Iterative construction of the partitions

Repeat this argument for layers $2, 3, \ldots, \ell - 1$, so within each part of $\mathcal{S}_{\ell-1}$, each of $f(x_1, \theta), \ldots, f(x_m, \theta)$ is a polynomial of degree ℓ in the p variables of θ.

$$2^m = \left|\left\{\left(\operatorname{sgn} f(x_i, \theta)\right)_i : \theta \in \mathbb{R}^p\right\}\right| \leq (6m\ell/p)^p \cdot |\mathcal{S}_{\ell-1}|$$

Iterative construction of the partitions

Repeat this argument for layers $2, 3, \ldots, \ell - 1$, so within each part of $\mathcal{S}_{\ell-1}$, each of $f(x_1, \theta), \ldots, f(x_m, \theta)$ is a polynomial of degree ℓ in the p variables of θ.

$$2^m = \left|\left\{(\operatorname{sgn} f(x_i, \theta))_i : \theta \in \mathbb{R}^p\right\}\right| \leq (6m\ell/p)^p \cdot |\mathcal{S}_{\ell-1}|$$

By construction of the partitions, $|\mathcal{S}_1| \leq (6m/k)^{k^2}$ and

$$\frac{|\mathcal{S}_n|}{|\mathcal{S}_{n-1}|} \leq \left(\frac{6m}{k}\right)^{nk^2} \qquad \text{for } n = 2, \ldots, \ell - 1, \text{ so}$$

$$2^m \leq \left(\frac{6m\ell}{p}\right)^p \prod_{n=1}^{\ell-1} \left(\frac{6m}{k}\right)^{nk^2} = \left(\frac{6m}{k}\right)^{\frac{k^2 \ell(\ell+1)}{2}} \leq m^{\ell p}, \text{ so}$$

$$m \leq O(\ell p \log(\ell p))$$

The upper bound

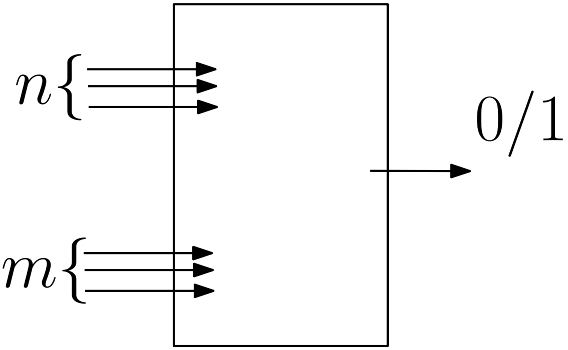
Theorem (Bartlett, Harvey, Liaw, M'17)

Let σ be a piecewise linear function, p be the number of parameters, ℓ the number of layers.
Then VC-dim $\leq O(p\ell \log p)$.

Next, lower bound:
there exist networks with VC-dim $\geq \Omega(p\ell \log(p/\ell))$.

Lower bound proof

Let $S_n := \{e_1, \ldots, e_n\}$ be standard basis for $\mathbb{R}^n$.
Build a network, input dimension $n + m$, shattering $S_n \times S_m$.
This implies VC-dimension $\geq |S_n \times S_m| = nm$



Lower bound proof

Let $S_n := \{e_1, \ldots, e_n\}$ be standard basis for $\mathbb{R}^n$.
Build a network, input dimension $n + m$, shattering $S_n \times S_m$.
This implies VC-dimension $\geq |S_n \times S_m| = nm$

For any given $g : S_n \times S_m \to \{0, 1\}$, find parameter vector θ such that $f((e_i, e_j), \theta) = g(e_i, e_j)$.

Lower bound proof

For any given $g : S_n \times S_m \to \{0, 1\}$, find parameter vector θ such that $f((e_i, e_j), p) = g(e_i, e_j)$.

Lower bound proof

For any given $g : S_n \times S_m \to \{0, 1\}$, find parameter vector θ such that $f((e_i, e_j), p) = g(e_i, e_j)$.

Build $n \times m$ table, entry $i, j = g(e_i, e_j)$:

$$\begin{vmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{vmatrix}$$

Lower bound proof

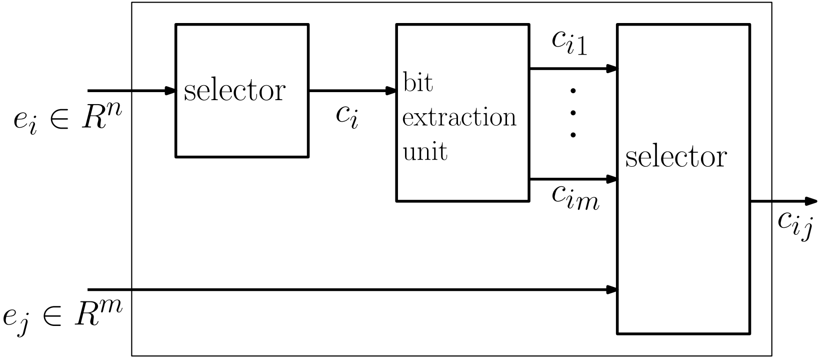
For any given $g : S_n \times S_m \to \{0,1\}$, find parameter vector θ such that $f((e_i, e_j), p) = g(e_i, e_j)$.

Build $n \times m$ table, entry $i, j = g(e_i, e_j)$:

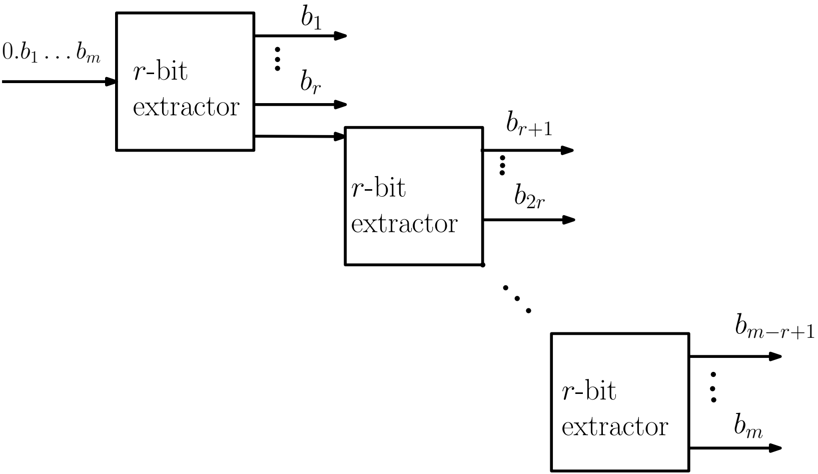
$$\begin{array}{r|ccc|} c_1 = 0. & 0 & 1 & 0 \\ c_2 = 0. & 1 & 1 & 0 \\ c_3 = 0. & 0 & 0 & 0 \\ c_4 = 0. & 1 & 1 & 0 \end{array}$$

Let $c_i \in [0,1]$ have binary representation the ith row.
On input (e_i, e_j), the network must output $g(e_i, e_j) = c_{ij}$, the jth bit of binary representation of c_i.

The bit extractor network

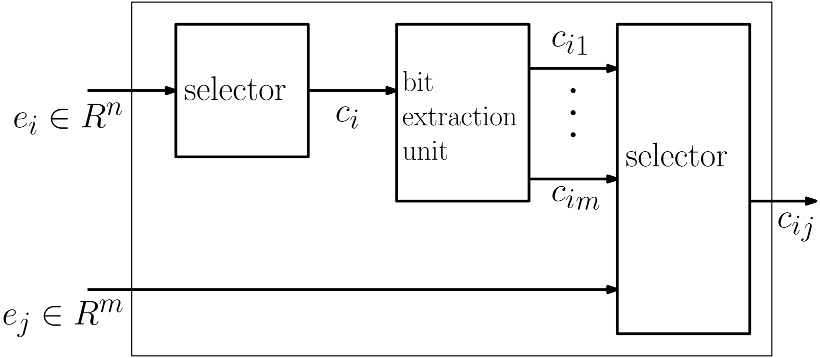


The bit extraction unit



Each block has 5 layers and $O(r2^r)$ parameters.
In total $O(m/r)$ layers and $O(m2^r)$ parameters..

The bit extractor network



Layers $= O(1 + m/r)$ and parameters $= O(n + m2^r + m)$

Given p, ℓ, let $r = \frac{1}{2}\log_2\left(\frac{p}{\ell}\right)$, $m = \frac{r\ell}{8}$, $n = \frac{p}{2}$

VC-dimension $\geq mn = \Omega(p\ell \log(p/\ell))$

Conclusion

Theorem (Vapnik and Chervonenkis 1971, Dudley 1978)

Given a training set of size m, expected estimation error is $\leq C\sqrt{\frac{\text{VC-dim}(H)}{m}}$.

Theorem (Bartlett, Harvey, Liaw, M, Conference on Learning Theory (COLT'17))

Let p be the number of parameters, ℓ the number of layers. Then for piecewise linear networks, VC-dim $\leq O(p\ell \log p)$ and there exist networks with VC-dim $\geq \Omega(p\ell \log(p/\ell))$.

These give a nearly-tight upper bound for the size of the training set needed to train a given neural network.

Further research on theory of neural networks

- ✓ Obtain tighter bounds by making additional assumptions, e.g. on the input data distribution.
- ✓ Understanding the optimization problem: how to choose the parameters to minimize the training error?
- ✓ How to design the network architecture for a given learning task?