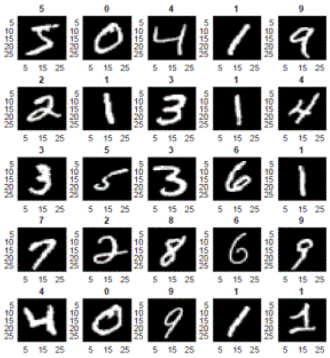
Goal: construct a network that can recognize handwritten numbers $\in \{0, 1, \dots, 9\}$
from MNIST (Modified National Institute of Standards and Technology) data set.



- contains 60000 training images, labeled by 'image ID' $n = 1, \dots, N$ ($= N$)
  and 10000 testing images
- 28x28 pixels, labeled by 'pixel ID' $l = 1, \dots, 784 \equiv L$
- each pixel contains grey-scale value $x^l_n \in (0, 1) \equiv I \subset R$
  white black unit interval
- image $n$ is represented by 'image vector' $\vec{x}_n = (x^1_n, \dots, x^L_n) \in I^L$
- each image has been assigned a 'target name' $\vec{t}_n \in \{\vec{e}_0, \dots, \vec{e}_9\}$,

  where $\vec{e}_j = (0, 0, 0, 1, 0, 0, 0)$, a basis vector in $N^{10}$, represents the number $j \in \{0, \dots, 9\}$
  ↑ position $j$

Goal: find 'decision function' $\vec{f}$ that maps image vector to 'predicted name',

$$\vec{f} : I^L \to N^{10}, \qquad \vec{x}_n \longmapsto \vec{f}(\vec{x}_n) \equiv \vec{f}_n$$

'predicted name'

while minimizing the cost function $$C = \sum_{n=1}^{N} (\vec{f}_n - \vec{t}_n)^2$$

target name

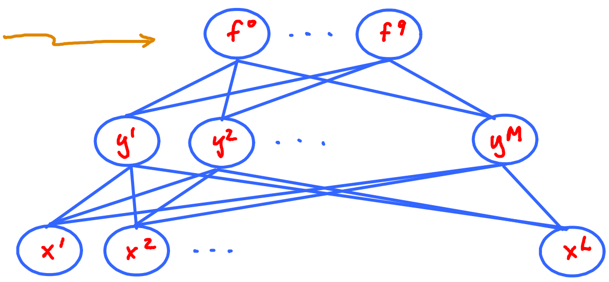
[Alternatively, choose $\vec{f} \in I^{10}$, $|\vec{f}| = 1$ then $f^j$ = probability that image is the number $j$ ]

## 1. Neural network

'output layer': $\vec{f} = (f^0, \dots, f^9) \in N^{10}$

'hidden layer': $\vec{y} = (y^1, \dots, y^M) \in I^L$

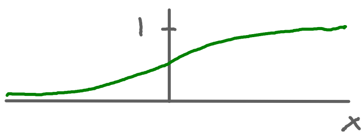
'input layer': $\vec{x} = (x^1, \dots, x^L) \in I^L$




Non-linear transformation: $y^k = \sigma(b^k + w^k_l x^l)$

'bias' 'weight' 'input'

with $\sigma(x) = \dfrac{1}{1 + e^{-x}}$ 'sigmoid function'




mimics neuron: 'fires' when input is above threshold 'soft-max layer': $f^j = \dfrac{e^{(a^j + u^j_\ell y^\ell)}}{\sum_{i=0}^{9} e^{(a^i + u^i_\ell y^\ell)}}$

use of exponentials emphasizes largest output at expense of others

$\vec{v} = (b, w, a, u)$ are variational parameters, used to minimize $C$ (e.g. by gradient descent)

$\Longrightarrow$ 'train the network' = 'supervised learning'

<u>Multilayer networks</u> (many layers = 'deep learning')

All of the above is just one possible Ansatz.
Many others can and have been tried.

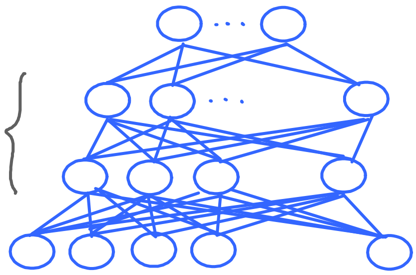
E.g.: multilayer networks:

hope is: will capture hierarchical structure better

output layer:

hidden layers:

input layer:



As before, sigmoid functions can be used to map input to output from one layer to the next.

Optimize cost function using gradient descent: $C = C(\vec{v})$

parameters of network

Gradient: $-\vec{\nabla} C = -\left(\frac{\partial C}{\partial v^1}, \frac{\partial C}{\partial v^2}, \dots \right)$ points in direction of steepest descent:

New variables: $\vec{v}\,' = \vec{v} - \eta \vec{\nabla} C$

'learning rate' (should be neither too small, nor too large)

## 2. Supervised learning with tensor networks

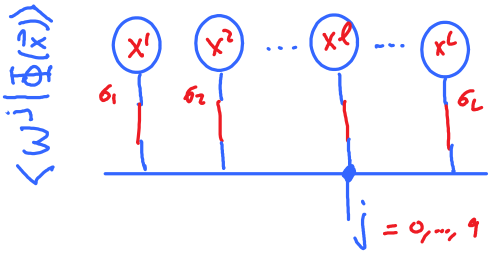
[Novikov2016], [Stoudenmire2017] with Schwab; [Maier2017] Bachelor thesis of David Maier

Goal: construct decision function $\vec{f}$ using a tensor network (here MPS);
train network using optimization techniques familiar from DMRG

Ansatz: $\vec{f} : I^L \longmapsto I^{10}$, (1)

$\vec{x} \longmapsto \vec{f}(\vec{x}) \equiv \langle \vec{W} | \Phi(\vec{x}) \rangle$ (2)

image vector predicted name




where right-hand side involves two separate maps:

'feature map' $\Phi : \vec{x} \mapsto |\Phi(\vec{x})\rangle$ : encodes greyscale input data into $L$ -leg MPS, $|\Phi(\vec{x})\rangle$ (3)

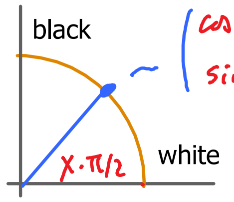
'weight vector' $\vec{W} : |\Phi(\vec{x})\rangle \longmapsto f^j(\vec{x}) \equiv \langle W^j | \Phi(\vec{x}) \rangle$, $j = 0, \dots, 9$ (4)

converts feature map into predicted name via inner product with an $L$-leg MPS,

'predicted name': that label $j$ for which $f^j$ is maximal.

### Feature map: encoding input data

map color range
(0,1) = (white, black)
to quarter-unit-circle,

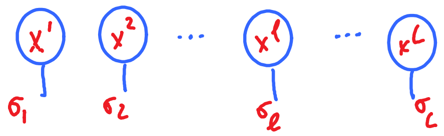



$\sim \begin{pmatrix} \cos \pi/2\, x \\ \sin \pi/2\, x \end{pmatrix} \equiv \begin{pmatrix} \phi_+(x) \\ \phi_-(x) \end{pmatrix} \equiv |\phi(x)\rangle$ (5)

so that $\langle \phi(x') | \phi(x) \rangle = \sum_{\sigma = \pm} \phi_\sigma(x') \phi_\sigma(x) \doteq \begin{cases} 1 & \text{if } x \simeq x' \\ 0 & \text{if } x \simeq \text{white}, \quad x' \simeq \text{black} \end{cases}$ (6)

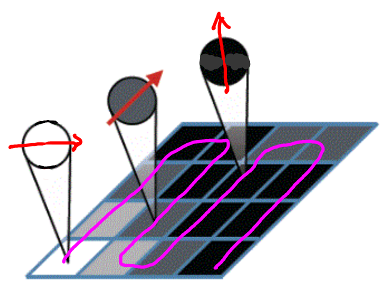
Choose 'snake-ordering' of pixels,
and encode image in a product state MPS:

$|\Phi(\vec{x})\rangle = |\phi(x^1)\rangle \otimes |\phi(x^2)\rangle \otimes \dots \otimes |\phi(x^L)\rangle$ (7)

$=$  (8)



$(d=2)$



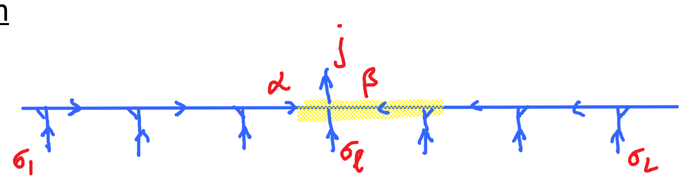
This construction for $|\Phi(\vec{x})\rangle$ is not unique. Other constructions are possible, provided that

$\langle \Phi(\vec{x}') | \Phi(\vec{x}) \rangle$ is a smooth and slowly varying function of $\vec{x}$ and $\vec{x}'$

which induces a 'distance matrix' in feature space which tends to cluster similar images together.

## Weight vector: encoding pattern recognition

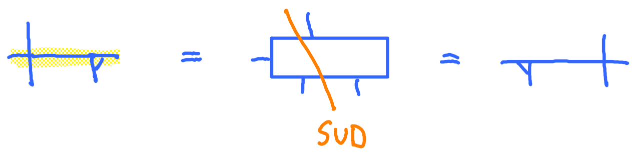
$|W^j\rangle$ = $L$-leg MPS =




$$= |\vec\sigma\rangle A^{\sigma_1} A^{\sigma_2} \dots M^{\sigma_\ell, j} \dots B^{\sigma_{L-1}} B^{\sigma_L}$$

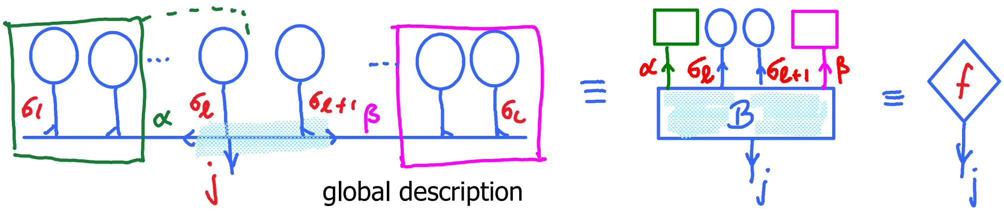
Left-normalized A's, right-normalized B's, sandwiching a 4-leg tensor, $M^{\alpha\sigma_\ell j \beta}$, at site $\ell$

Top leg can be moved around:

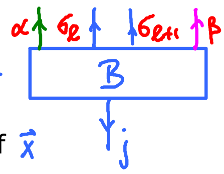



Decision function:

$$f^j(\vec x) \overset{(4)}{=} \langle W^j | \Phi(\vec x)\rangle =$$

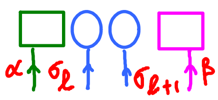



$\equiv \langle B^j | \tilde\Phi(\vec x)\rangle$, with $\langle B^j | =$




independent of $\vec x$

, $|\tilde\Phi(\vec x)\rangle =$



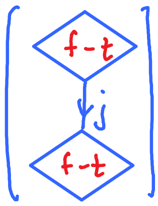

independent of $j$

Note: all $\vec x$-dependence resides in $|\tilde\Phi(\vec x)\rangle$, all $j$-dependence in $|B^j\rangle$.

Location of 'central site' can be shifted (e.g. during sweeping).

## Cost function

$$C = \sum_{n=1}^{N} (\vec f_n - \vec t_n)^2 = \sum_{n=1}^{N}$$




, $\vec f_n \equiv \vec f(\vec x_n)$

evaluated at $\vec x_n$

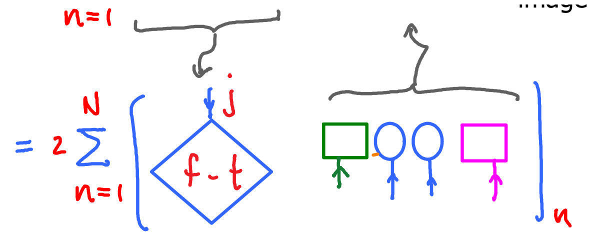
For given set of training data $\{\vec x_n, \vec t_n \,|\, n = 1, \dots, N\}$, minimize $C$ w.r.t. $\langle \vec W|$, or equivalently, $\langle \vec B|$.

Minimize using gradient steepest descent. Compute the gradient:

$$|\nabla B^j\rangle^{\alpha\sigma_\ell\sigma_{\ell+1}\beta} \equiv \frac{\partial C}{\partial B^j} \equiv \sum_{n=1}^{N} (f_n^j - t_n^j) \frac{\partial f_n^j}{\partial B^j} = 2\sum_{n=1}^{N} (f_n^j - t_n^j) \, |\hat\Phi(\vec x_n)\rangle$$

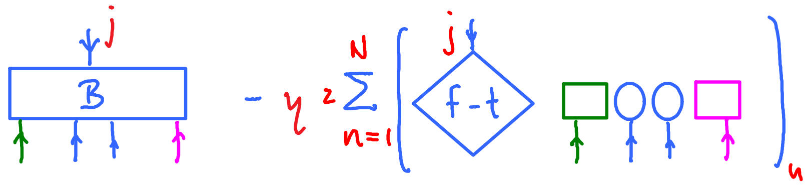
sum over training set

differs from one image to the next



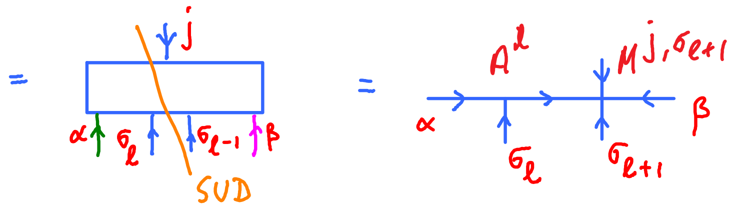
Then update the MPS:

$$|B'^j\rangle = |B^j\rangle - \eta\, |\nabla B^j\rangle =$$

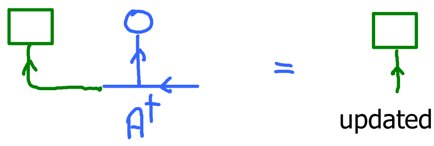


learning rate
(must be chosen very carefully!)

Advance to next site:




Update training input:




Sweep back and forth until A-tensors no longer change -- then 'training of network' is complete.

<u>Comments</u>

Costs: $\mathcal{O}(d^3 D^3 N \cdot L \cdot 10)$

$d$ : physical bond dimension (here: $2$ )

$N$ : number of training images

$D$ : MPS bond dimension (free parameter)

$L$ : number of pixels per image

Once network has been trained, prediction of a new image $\vec{x}$ proceeds simply via

$$f^j(\vec{x}) = \langle W^j | \Phi(\vec{x}) \rangle$$

, predicted name is the $j$ yielding maximal $f^j$

MNIST test:

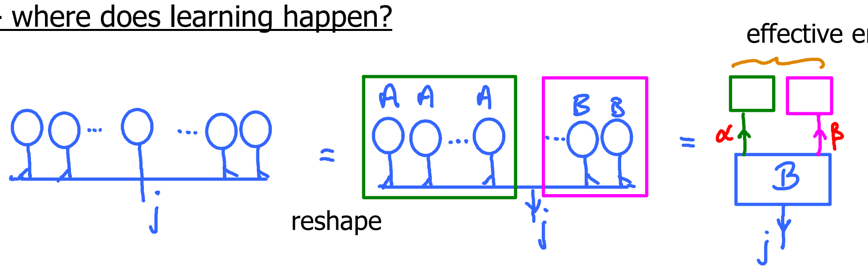
- 28 x 28 was coarse-grained to 14 x 14 (to save resources)
- at most 5 sweeps were needed before training converges
- bond dimension $D = 10 \Rightarrow$ 5% error rate

  $20 \Rightarrow$ 2% error rate

  $120 \Rightarrow$ 0.97% error rate

## Implicit feature selection - where does learning happen?

$f^j = \langle W^j | \Phi \rangle =$




$= \langle \tilde{\tilde{B}}^j | \tilde{\tilde{\Phi}} \rangle$

- $|\tilde{\tilde{\Phi}}\rangle$ is projection of $|\Phi\rangle$ onto space spanned by orthonormal basis, encoded in $\langle \tilde{\tilde{B}}^j|$

  has just $D^2$ components

- So, training an MPS model uncovers relatively small set of features, and simultaneously trains decision function using only those features.

- 'Feature selection' occurs when computing SVD: basis elements which do not contribute optimally to bond tensors are discarded

## Future prospects

- try tensor networks that are designed for 2D (PEPS, TRG, MERA,)
- exploit symmetries
- try other sampling schemes
- ~~try other sampling schemes~~
- 'unsupervised learning' with tensor networks
- ...