Neuroscience-inspired online unsupervised learning algorithms

Presented by Lili Su, Jiajia Zhao

CSAIL, MIT

April 24, 2020

Neuroscience-Inspired Learning: Motivation

Deep learning:

- Has superior performance in practice
  - applications: computer vision, speech recognition, natural language processing, audio recognition, etc.
- Was derived their inspiration from biology

Neuroscience-Inspired Learning: Motivation

Deep learning:

- Has superior performance in practice
  - applications: computer vision, speech recognition, natural language processing, audio recognition, etc.
- Was derived their inspiration from biology

"Drawbacks" compared with the brain:

- Vulnerable to adversarial noises
- High energy consumption
- Hardware requirements

Neuroscience-Inspired Learning: Motivation

Deep learning:

- Has superior performance in practice
  - applications: computer vision, speech recognition, natural language processing, audio recognition, etc.
- Was derived their inspiration from biology

"Drawbacks" compared with the brain:

- Vulnerable to adversarial noises
- High energy consumption
- Hardware requirements

Further inspiration from the brain might be helpful

Focus of This Paper: Bio-Plausibility

Most artificial NNs resemble natural NNs only superficially

- Training (synaptic strength update) is not bio-plausible
  – backpropagation

Focus here: Bio-plausibility on the algorithmic level

- not attempt to reproduce many biological details (not ion channels)
- develop algorithms that respect major biological constraints

Bio-Plausibility: Be "Online" and Be "Local"

- (Be "Online"): input data are streamed to the algorithm (neural circuits) sequentially, and the corresponding output must be computed before the next input sample arrives
- (Be "Local"): a biological synapse can update its weight based on the activity of only the two neurons that the synapse connects.
  - Such "locality" of the learning rule is violated by most artificial NNs including backpropagation-based deep learning networks.

Key Contributions-I

A family of biologically plausible artificial neural networks (NNs) for unsupervised learning

- The inspiration from the brain:
  Signal processing in the brain tends to preserves similarity
  [Qin-Mudur-Pehlevan'20]
- Mathematically:
  - Use a family of *principled* objective functions containing a term that penalizes dissimilarity
  - Derive the NNs by running *alternating stochastic gradient descent* on the corresponding objective functions

Key Contributions (Continued)

This family of objective functions cover a large range of interesting machine learning problems, such as

1. linear dimensionality reduction (PCA);
2. sparse and/or nonnegative feature extraction;
3. blind nonnegative source separation;
4. clustering and manifold learning.

Key Contributions (Continued)

This family of objective functions cover a large range of interesting machine learning problems, such as

1. **linear dimensionality reduction (PCA)**;
2. sparse and/or nonnegative feature extraction;
3. blind nonnegative source separation;
4. clustering and manifold learning.

Linear dimensionality reduction (PCA)

Outline of the remainder

1. Background: Why is the previous work not satisfactory
    - Extending Oja's rule to multiple output neurons setting

Outline of the remainder

1. Background: Why is the previous work not satisfactory
    - Extending Oja's rule to multiple output neurons setting
    - Changing the objective might suffice?

Outline of the remainder

1. Background: Why is the previous work not satisfactory
    - Extending Oja's rule to multiple output neurons setting
    - Changing the objective might suffice?
2. Similarity-based approach
    - Similarity-based objectives
    - Local learning rules obtained by alternating stochastic gradient descent
        - Key technique: Variable substitution trick

Outline of the remainder

1. Background: Why is the previous work not satisfactory
   - Extending Oja's rule to multiple output neurons setting
   - Changing the objective might suffice?
2. Similarity-based approach
   - Similarity-based objectives
   - Local learning rules obtained by alternating stochastic gradient descent
     - Key technique: Variable substitution trick
3. Beyond PCA: Other tasks solved by the similarity-based approach

Outline of the remainder

1. Background: Why is the previous work not satisfactory
   - Extending Oja's rule to multiple output neurons setting
   - Changing the objective might suffice?
2. Similarity-based approach
   - Similarity-based objectives
   - Local learning rules obtained by alternating stochastic gradient descent
     - Key technique: Variable substitution trick
3. Beyond PCA: Other tasks solved by the similarity-based approach
4. Summary and conclusions

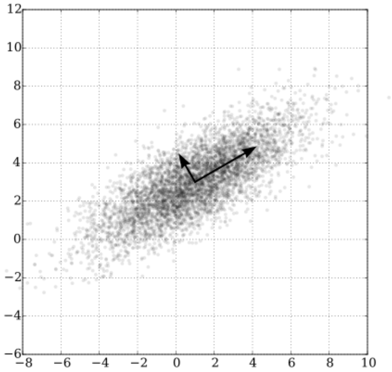
Principal Component Analysis (PCA)

PCA in plain words:

- The first component is the a "best fitting" **line**
- The second component is the next best-fitting line and is perpendicular to the first

In general,

- subspace projection
- orthogonality of the components



Principal Component Analysis (PCA)

Mathematically:

- Given T data points $\{x_1, \cdots, x_T\} \subseteq \mathbb{R}^n$
- Let $u \in \mathbb{R}^n$ such that $\|u\|_2 = 1$.
- The first/top principle of the given dataset is

$$u^* = \arg\min_u \frac{1}{2T} \sum_{i=1}^{T} \|x_i - \langle x_i, u \rangle u\|_2^2.$$

Oja's rule can be viewed as running SGD on the above objective

Oja's rule finds the first principle

Connection between SGD v.s. Oja's Rule

PCA: $u^* = \arg\min_u \frac{1}{2T} \sum_{i=1}^{T} \|x_i - \langle x_i, u \rangle u\|_2^2$

rewriting $\|x_i - \langle x_i, u \rangle u\|_2^2 = \min_y \|x_i - yu\|_2^2$

Derivation of Oja's Rule:

- Update u_t via SGD: Let $y_t = \langle x_t, u_{t-1} \rangle$

$$u_t \leftarrow \frac{u_{t-1} + \eta \langle x_t, u_{t-1} \rangle x_t}{\|u_{t-1} + \eta \langle x_t, u_{t-1} \rangle x_t\|_2} = u_{t-1} + \eta(x_t - u_{t-1} y_t) y_t + O(\eta^2)$$

Connection between SGD v.s. Oja's Rule

PCA: $u^* = \arg\min_u \frac{1}{2T} \sum_{i=1}^{T} \|x_i - \langle x_i, u \rangle u\|_2^2$

rewriting $\|x_i - \langle x_i, u \rangle u\|_2^2 = \min_y \|x_i - yu\|_2^2$

Derivation of Oja's Rule:

- Update u_t via SGD: Let $y_t = \langle x_t, u_{t-1} \rangle$

$$u_t \leftarrow \frac{u_{t-1} + \eta \langle x_t, u_{t-1} \rangle x_t}{\|u_{t-1} + \eta \langle x_t, u_{t-1} \rangle x_t\|_2} = u_{t-1} + \eta(x_t - u_{t-1} y_t) y_t + O(\eta^2)$$

- Update u_t via alternating SGD: First minimize y, then SGD on u

$$y_t \leftarrow \langle x_t, u_{t-1} \rangle \text{; and } u_t \leftarrow u_{t-1} + \eta(x_t - u_{t-1} y_t) y_t.$$

Connection between SGD v.s. Oja's Rule

PCA: $u^* = \arg\min_u \frac{1}{2T} \sum_{i=1}^{T} \|x_i - \langle x_i, u\rangle\, u\|_2^2$

rewriting $\|x_i - \langle x_i, u\rangle\, u\|_2^2 = \min_y \|x_i - yu\|_2^2$

Derivation of Oja's Rule:

- Update u_t via SGD: Let $y_t = \langle x_t, u_{t-1}\rangle$

$$u_t \leftarrow \frac{u_{t-1} + \eta\, \langle x_t, u_{t-1}\rangle\, x_t}{\|u_{t-1} + \eta\, \langle x_t, u_{t-1}\rangle\, x_t\|_2} = u_{t-1} + \eta(x_t - u_{t-1}y_t)y_t + O(\eta^2)$$

- Update u_t via alternating SGD: First minimize y, then SGD on u

$$y_t \leftarrow \langle x_t, u_{t-1}\rangle \,; \text{and } u_t \leftarrow u_{t-1} + \eta(x_t - u_{t-1}y_t)y_t.$$

Oja's rule finds the first component!!!

How about the top k components?

How about The Top k Components?

Question

Is it possible to solve the online general PCA algorithms using multiple neurons with bio-plausible updates?

- The objective for top component

$$u^* = \arg \min_{u:\|u\|_2=1} \frac{1}{2T} \sum_{i=1}^{T} \min_{y_i} \|x_i - y_i u\|_2^2$$

- The objective for top k components

$$W^* = \arg \min_{W:W\in\mathbb{R}^{n\times k}} \frac{1}{2T} \sum_{i=1}^{T} \min_{y_i} \|x_i - W y_i\|_2^2$$

How about The Top k Components?

Question

Is it possible to solve the online general PCA algorithms using multiple neurons with bio-plausible updates?

- The objective for top component

$$u^* = \arg \min_{u:\|u\|_2=1} \frac{1}{2T} \sum_{i=1}^{T} \min_{y_i} \|x_i - y_i u\|_2^2$$

- The objective for top k components

$$W^* = \arg \min_{W:W\in\mathbb{R}^{n\times k}} \frac{1}{2T} \sum_{i=1}^{T} \min_{y_i} \|x_i - Wy_i\|_2^2$$

Alternating SGD on the blue-colored objective is no longer bio-plausible

Alternating SGD is NO LONGER Bio-Plausible

$$W^* = \arg\min_{W: W \in \mathbb{R}^{n\times k}} \frac{1}{2T} \sum_{i=1}^{T} \min_{y_i} \|x_i - W y_i\|_2^2$$

- Let y_t^ℓ be the ℓ-th entry of y_t.
- Let W_{t-1}^j be the j–th column of W_{t-1} and W_{t-1}^{ij} be the entry at the i-th row and the j–th column.
- Update of y: $\nabla y_t^\ell = \left\langle W_{t-1}^\ell, x_t \right\rangle - \sum_{j=1}^{k} \left\langle W_{t-1}^\ell, W_{t-1}^j \right\rangle y_t^j$
- Update of W: $W_t^{ij} = W_{t-1}^{ij} + \eta \left(x_t^i - \sum_{\ell=1}^{k} W_{t-1}^{\ell i} y_t^\ell \right) y_t^j$

Alternating SGD is NO LONGER Bio-Plausible

$$W^* = \arg\min_{W: W \in \mathbb{R}^{n \times k}} \frac{1}{2T} \sum_{i=1}^{T} \min_{y_i} \|x_i - W y_i\|_2^2$$

- Let y_t^ℓ be the ℓ-th entry of y_t.
- Let W_{t-1}^j be the j–th column of W_{t-1} and W_{t-1}^{ij} be the entry at the i-th row and the j–th column.
- Update of y: $\nabla y_t^\ell = \left\langle W_{t-1}^\ell, x_t \right\rangle - \sum_{j=1}^k \left\langle W_{t-1}^\ell, W_{t-1}^j \right\rangle y_t^j$
- Update of W: $W_t^{ij} = W_{t-1}^{ij} + \eta \left(x_t^i - \sum_{\ell=1}^k W_{t-1}^{\ell i} y_t^\ell \right) y_t^j$

Is it possible to find an alternative objective for PCA?

Similarity-based Objective Function

$$\min_{y_1,\ldots y_T} \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} ({x_t}^\top x_{t'} - {y_t}^\top y_{t'})^2$$

- Similarity: dot product for a pair of inputs ($\mathbb{R}^n$) or outputs ($\mathbb{R}^k, k < n$).

Similarity-based Objective Function

$$\min_{y_1,\ldots y_T} \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} (x_t{}^\top x_{t'} - y_t{}^\top y_{t'})^2$$

- Similarity: dot product for a pair of inputs ($\mathbb{R}^n$) or outputs ($\mathbb{R}^k, k < n$).
- Matching: want similarity of inputs and that of outputs to be close

Similarity-based Objective Function

$$\min_{y_1,\dots y_T} \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} (x_t{}^\top x_{t'} - y_t{}^\top y_{t'})^2$$

- Similarity: dot product for a pair of inputs ($\mathbb{R}^n$) or outputs ($\mathbb{R}^k, k < n$).
- Matching: want similarity of inputs and that of outputs to be close
- Offline solution: unique global PCA
  solution up to an orthogonal rotation, aka principal subspace projection

[Pehlevan-Chklovski, NeurIPS 1

Problems Going Online?

$$\min_{y_1,\ldots y_T} \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} ({x_t}^\top x_{t'} - {y_t}^\top y_{t'})^2$$

- Require information from other time steps (non-online and non-local)

Problems Going Online?

$$\min_{y_1,\dots y_T} \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} ({x_t}^\top x_{t'} - {y_t}^\top y_{t'})^2$$

- Require information from other time steps (non-online and non-local)
- Mapping onto NN with synaptic weight updates (unclear if bio-plausible)

Problems Going Online?

$$\min_{y_1,\dots y_T} \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}({x_t}^\top x_{t'} - {y_t}^\top y_{t'})^2$$

- Require information from other time steps (non-online and non-local)
- Mapping onto NN with synaptic weight updates (unclear if bio-plausible)

$$\begin{aligned}
&\min_{y_1,\dots y_T} \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}({x_t}^\top x_{t'} - {y_t}^\top y_{t'})^2 \\
&= \min_{y_1,\dots y_T} \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}(-2y_t^\top y_{t'} x_t^\top x_{t'} + y_t^\top y_{t'} y_t^\top y_{t'}) \\
&= \min_{y_1,\dots y_T} -\frac{2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} + \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} y_t^\top y_{t'}
\end{aligned}$$

Variable Substitution Trick

Key intuition: for any given a, $(a-b)^2 \geq 0$ is minimized when $b = a$.

Variable Substitution Trick

Key intuition: for any given a, $(a-b)^2 \geq 0$ is minimized when $b = a$.
Apply to Matrix W:

$$
\begin{aligned}
0 &\leq \left\langle W - \frac{1}{T}\sum_{t=1}^{T} x_t y_t^\top, W - \frac{1}{T}\sum_{t=1}^{T} x_t y_t^\top \right\rangle \\
&= \operatorname{Tr} W^\top W - \frac{2}{T}\sum_{t=1}^{T} y_t^\top W x_t + \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} \\
&=> -\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} \leq \operatorname{Tr} W^\top W - \frac{2}{T}\sum_{t=1}^{T} y_t^\top W x_t
\end{aligned}
$$

Variable Substitution Trick

Key intuition: for any given a, $(a-b)^2 \geq 0$ is minimized when $b = a$.
Apply to Matrix W:

$$0 \leq \left\langle W - \frac{1}{T}\sum_{t=1}^{T} x_t y_t^\top, W - \frac{1}{T}\sum_{t=1}^{T} x_t y_t^\top \right\rangle$$

$$= \operatorname{Tr} W^\top W - \frac{2}{T}\sum_{t=1}^{T} y_t^\top W x_t + \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'}$$

$$=> -\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} \leq \operatorname{Tr} W^\top W - \frac{2}{T}\sum_{t=1}^{T} y_t^\top W x_t$$

where the equality holds iff $W = \frac{1}{T}\sum_{t=1}^{T} y_t x_t^\top$, i.e.

$$-\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} = \min_{W\in\mathbb{R}^{k\times n}} \operatorname{Tr} W^\top W - \frac{2}{T}\sum_{t=1}^{T} y_t^\top W x_t.$$

Variable Substitution Trick

Objective func:

$$\min_{y_1,\dots y_T} -\frac{2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} + \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} y_t^\top y_{t'}$$

Variable Substitution Trick

Objective func:

$$\min_{y_1,\dots y_T} -\frac{2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} + \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} y_t^\top y_{t'}$$

First term:

$$-\frac{2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} = \min_{W\in\mathbb{R}^{k\times n}} 2\operatorname{Tr} W^\top W - \frac{4}{T}\sum_{t=1}^{T} y_t^\top W x_t .$$

Variable Substitution Trick

Objective func:

$$\min_{y_1,\dots y_T} -\frac{2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} + \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} y_t^\top y_{t'}$$

First term:

$$-\frac{2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} = \min_{W\in\mathbb{R}^{k\times n}} 2\,\mathrm{Tr}\,W^\top W - \frac{4}{T}\sum_{t=1}^{T} y_t^\top W x_t.$$

Similarly,

$$\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} y_t^\top y_{t'} = \max_{M\in\mathbb{R}^{k\times k}} \frac{2}{T}\sum_{t=1}^{T} y_t^\top M y_t - \mathrm{Tr}\,M^\top M$$

Variable Substitution Trick

Objective func:

$$\min_{y_1,\dots y_T} -\frac{2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} + \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} y_t^\top y_{t'}$$

First term:

$$-\frac{2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} = \min_{W\in\mathbb{R}^{k\times n}} 2\,\mathrm{Tr}\, W^\top W - \frac{4}{T}\sum_{t=1}^{T} y_t^\top W x_t.$$

Similarly,

$$\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} y_t^\top y_{t'} = \max_{M\in\mathbb{R}^{k\times k}} \frac{2}{T}\sum_{t=1}^{T} y_t^\top M y_t - \mathrm{Tr}\, M^\top M$$

New form of the objective function is local in the online setting!

New Form of Objective Function

$$\min_{y_1,\dots y_T} \left[\min_{W\in\mathbb{R}^{k\times n}} \max_{M\in\mathbb{R}^{k\times k}} \frac{1}{T}\sum_{t=1}^{T}\left[2\,\mathrm{Tr}\,W^\top W - \mathrm{Tr}\,M^\top M + l_t(W,M,y_t)\right]\right]$$

$$= \min_{W\in\mathbb{R}^{k\times n}} \max_{M\in\mathbb{R}^{k\times k}} \frac{1}{T}\sum_{t=1}^{T}\left[2\,\mathrm{Tr}\,W^\top W - \mathrm{Tr}\,M^\top M + \min_{y_t} l_t(W,M,y_t)\right]$$

where

$$l_t(W,M,y_t) = -4x_t^\top W^\top y_t + 2y_t^\top M y_t$$

We have successfully separated the computations of outputs at different time steps, satisfying the requirement to be local.

Gradient Descent/Ascent Algorithm

- Gradient descent (minimizing) $l_t(W, M, y_t)$ wrt y_t

$$\frac{d}{d_{y_t}}(-4x_t^\top W^\top y_t + 2y_t^\top My_t) = -4(Wx_t - My_t)$$
$$=> \dot{y}_t = Wx_t - My_t$$

Gradient Descent/Ascent Algorithm

- Gradient descent (minimizing) $l_t(W, M, y_t)$ wrt y_t

$$\frac{d}{d_{y_t}}(-4x_t^\top W^\top y_t + 2y_t^\top My_t) = -4(Wx_t - My_t)$$
$$=> \dot{y}_t = Wx_t - My_t$$

- Gradient descent (minimizing) objective function wrt W

$$W_{ij} \leftarrow W_{ij} + \eta(y_i x_j - W_{ij})$$

Gradient Descent/Ascent Algorithm

- Gradient descent (minimizing) $l_t(W, M, y_t)$ wrt y_t

$$\frac{d}{d_{y_t}}(-4x_t^\top W^\top y_t + 2y_t^\top My_t) = -4(Wx_t - My_t)$$
$$=> \dot{y}_t = Wx_t - My_t$$

- Gradient descent (minimizing) objective function wrt W

$$W_{ij} \leftarrow W_{ij} + \eta(y_i x_j - W_{ij})$$

- Gradient ascent (maximizing) objective function wrt M

$$M_{ij} \leftarrow M_{ij} + \frac{\eta}{2}(y_i y_j - M_{ij})$$

Gradient Descent/Ascent Algorithm

- Gradient descent (minimizing) $l_t(W, M, y_t)$ wrt y_t

$$\frac{d}{d_{y_t}}(-4x_t^\top W^\top y_t + 2y_t^\top M y_t) = -4(Wx_t - My_t)$$
$$=> \dot{y}_t = Wx_t - My_t$$

- Gradient descent (minimizing) objective function wrt W

$$W_{ij} \leftarrow W_{ij} + \eta(y_i x_j - W_{ij})$$

- Gradient ascent (maximizing) objective function wrt M

$$M_{ij} \leftarrow M_{ij} + \frac{\eta}{2}(y_i y_j - M_{ij})$$

Interpretation

- W and M represent synaptic weight changes in feed-forward and lateral connections. W and -M correspond to Hebbian/Anti-Hebbian.

Remaining Issues

There are still several issues remaining...

- Not exactly recovering principal components, but principal subspace projections

Remaining Issues

There are still several issues remaining...

- Not exactly recovering principal components, but principal subspace projections
- Recurrent activity on output neurons must settle faster than input variations

Remaining Issues

There are still several issues remaining...

- Not exactly recovering principal components, but principal subspace projections
- Recurrent activity on output neurons must settle faster than input variations
- Output neurons compete with each other with lateral connections - in real brains, have to go through inter-neurons

Whitening Constraints and Inter-neurons

- In order to derive PCA algorithms, change the objective function to encourage orthogonality of W (1).

Whitening Constraints and Inter-neurons

- In order to derive PCA algorithms, change the objective function to encourage orthogonality of W (1).
- Replace M with a whitening constraint (2):

$$\min_{y_1,\dots y_T} -\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'}, s.t. \frac{1}{T}\sum_t y_t y_t^\top = I_k$$

where I_k is the k-by-k identity matrix.

Whitening Constraints and Inter-neurons

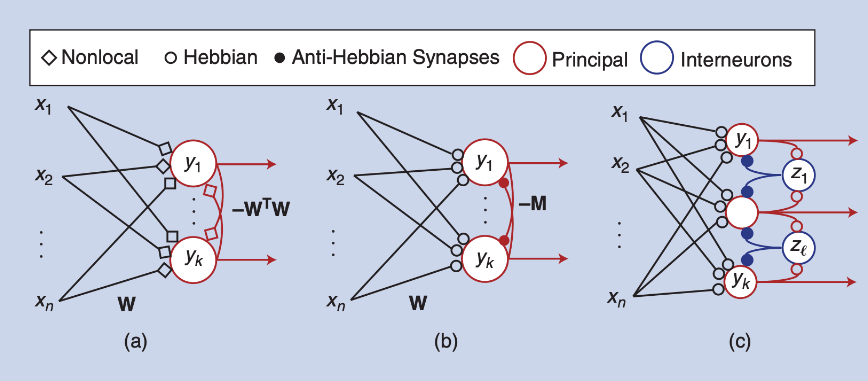
- In order to derive PCA algorithms, change the objective function to encourage orthogonality of W (1).
- Replace M with a whitening constraint (2):

$$\min_{y_1,\dots y_T} -\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'}, s.t. \frac{1}{T}\sum_t y_t y_t^\top = I_k$$

  where I_k is the k-by-k identity matrix.
- Objective function modeled by Lagrange formalism:

$$\min_{y_1,\dots y_T}\max_{z_1,\dots,z_T} -\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} + \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} z_t^\top z_{t'} (y_t^\top y_{t'} - \delta_{t,t'})$$

where $\delta_{t,t'}$ is the Kronecker delta, and $z_t^\top z_{t'}$ naturally model interneuron activities. See details in (3).



◇ Nonlocal
○ Hebbian
● Anti-Hebbian Synapses
Principal
Interneurons
x_1
x_2
x_n
y_1
y_k
$-\mathbf{W}^T\mathbf{W}$
W
(a)
x_1
x_2
x_n
y_1
y_k
$-\mathbf{M}$
W
(b)
x_1
x_2
x_n
y_1
y_k
z_1
z_ℓ
(c)

Beyond PCA:

Other tasks solved by the similarity-based approach

Nonnegative Similarity-Matching Objective

- "Nonnegative": variable constraints – this constraint corresponds to ReLU activation

- The minimization problem becomes

$$\min_{y_1,\ldots y_T \geq 0} \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} ({x_t}^\top x_{t'} - {y_t}^\top y_{t'})^2 \tag{1}$$

Nonnegative Similarity-Matching Objective

- "Nonnegative": variable constraints – this constraint corresponds to ReLU activation
- The minimization problem becomes

$$\min_{y_1,\ldots y_T \geq 0} \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} ({x_t}^\top x_{t'} - {y_t}^\top y_{t'})^2 \tag{1}$$

Equation (1) can be solved by the same learning rule **except that**

- the output variables are projected onto the nonnegative domain

Visualization

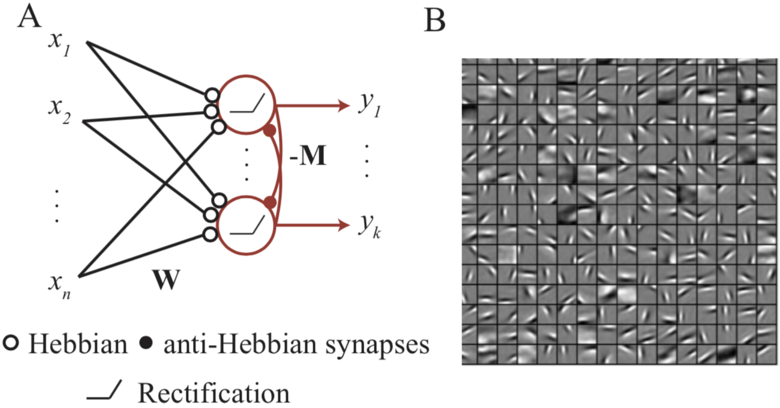


Figure: B) Nonnegative similarity matching learns edge filters from patches of whitened natural scenes.

Nonnegative Similarity-Matching for Clustering

K-means clustering (MacQueen, 1967)

Let $\mathcal{C}_1, \cdots, \mathcal{C}_K$ be a partition of the T data points $x_1, \cdots, x_T$. Want to find a best partition such that

$$\min_{\mathcal{C}_1, \cdots, \mathcal{C}_K} \sum_{k=1}^{K} \sum_{t \in \mathcal{C}_k} \left\| x_t - \frac{1}{n_k} \sum_{s \in \mathcal{C}_k} x_s \right\|_2^2, \qquad \text{where } n_k := |\mathcal{C}_k|$$

Let $Y \in \mathbb{R}^{K \times T}$ be a scaled indicator matrix s.t.

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_{:,1} \\ \vdots \\ \boldsymbol{y}_{:,K} \end{pmatrix}, \quad \boldsymbol{y}_{:,k} = \frac{1}{n_k^{1/2}} \left(0, \cdots, 0, \overbrace{1, \cdots, 1}^{n_k}, 0, \cdots, 0 \right).$$

Nonnegative Similarity-Matching for Clustering

K-means clustering (MacQueen, 1967)

Let $\mathcal{C}_1, \cdots, \mathcal{C}_K$ be a partition of the T data points $x_1, \cdots, x_T$. Want to find a best partition such that

$$\min_{\mathcal{C}_1,\cdots,\mathcal{C}_K} \sum_{k=1}^{K} \sum_{t\in\mathcal{C}_k} \left\| x_t - \frac{1}{n_k} \sum_{s\in\mathcal{C}_k} x_s \right\|_2^2, \qquad \text{where } n_k := |\mathcal{C}_k|$$

Let $Y \in \mathbb{R}^{K\times T}$ be a scaled indicator matrix s.t.

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_{:,1} \\ \vdots \\ \boldsymbol{y}_{:,K} \end{pmatrix}, \quad \boldsymbol{y}_{:,k} = \frac{1}{n_k^{1/2}} \left(0,\cdots,0,\overbrace{1,\cdots,1}^{n_k},0,\cdots,0 \right).$$

$Y* = \arg\min_Y \|X^\top X - Y^\top Y\|$ gives a optimal K-means clustering

A Simplified Objective for Soft-Clustering

Why a simplified objective?

Finding the optimal solution is rather challenging factorization problem.

- The simplified objective is chosen so that clustering of inputs is based on input pairwise similarities

A Simplified Objective for Soft-Clustering

Why a simplified objective?

Finding the optimal solution is rather challenging factorization problem.

- The simplified objective is chosen so that clustering of inputs is based on input pairwise similarities

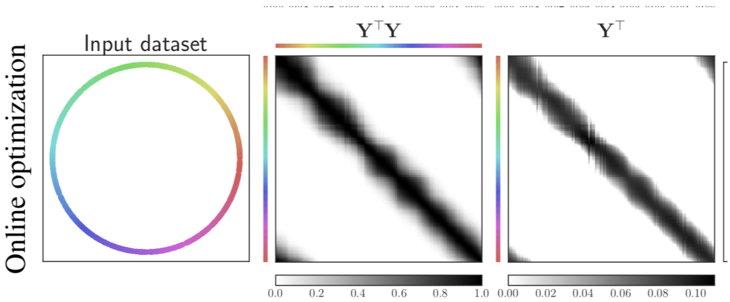
$$\min_{y_1,\ldots y_T \geq 0} \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} \left(\alpha - x_t^\top x_{t'}\right) y_t^\top y_{t'}$$
$$\text{s.t.} \quad \|y_t\|_2 \leq 1 \quad \forall t$$

  Here $\alpha > 0$ is the clustering threshold.

- Correctness of the clustering algorithm depends on how well the inputs are separated!!!

Manifold Tiling

In many real-world problems, data points are not well-segregated but lie on low-dimensional manifolds. For such data, the optimal solution of the above simplified objective effectively tiles the data manifold



Conclusions

$$\min_{\forall t, y_t \in \Omega} \left[- \sum_{t=1}^{T} \sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} + f(y_1, \ldots, y_T) \right]$$

Conclusions

$$\min_{\forall t, y_t \in \Omega} \left[- \sum_{t=1}^{T} \sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} + f(y_1, ..., y_T) \right]$$

- First term: the covariance of the similarity of the outputs and that of inputs.

Conclusions

$$\min_{\forall t, y_t \in \Omega} \left[-\sum_{t=1}^{T} \sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} + f(y_1, ..., y_T) \right]$$

- First term: the covariance of the similarity of the outputs and that of inputs.
- Optimizing first term online gives rise to synaptic local learning rules.

Conclusions

$$\min_{\forall t, y_t \in \Omega} \left[-\sum_{t=1}^{T} \sum_{t'=1}^{T} y_t^\top y_{t'} x_t^\top x_{t'} + f(y_1, ..., y_T) \right]$$

- First term: the covariance of the similarity of the outputs and that of inputs.
- Optimizing first term online gives rise to synaptic local learning rules.
- Second term f and constraints Ω: inhibitory mechanisms and other constraints that make the NN bio-plausible.

Optimization Feature	Biological Feature
Similarity (anti)alignment	(Anti-)Hebbian plasticity [16], [17]
Nonnegativity constraint	Rectifying neuron activation function [18], [21]
Sparsity regularizer	Adaptive neural threshold [40]
Constrained output correlation matrix	Adaptive lateral weights [7], [18]
Constrained PSD output Gramian	Anti-Hebbian interneurons [7]
Copositive output Gramian	Anti-Hebbian inhibitory neurons [31]
Constrained activity l_1-norm	Giant interneuron [36]

Future Work

Future Work

- Convergence Proof of online algorithms

Future Work

- Convergence Proof of online algorithms
- Supervised/Semi-supervised learning with reinforcement

Future Work

- Convergence Proof of online algorithms
- Supervised/Semi-supervised learning with reinforcement
- Temporal correlations in time in input data points

Future Work

- Convergence Proof of online algorithms
- Supervised/Semi-supervised learning with reinforcement
- Temporal correlations in time in input data points
- Stacking similarity-based NNs

Future Work

- Convergence Proof of online algorithms
- Supervised/Semi-supervised learning with reinforcement
- Temporal correlations in time in input data points
- Stacking similarity-based NNs
- Spikes in biological NNs - all or nothing

[1] C. Pehlevan and D. B. Chklovskii, "Optimization theory of hebbian/anti-hebbian networks for pca and whitening," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1458–1465, 2015.

[2] V. Minden, C. Pehlevan, and D. B. Chklovskii, "Biologically plausible online principal component analysis without recurrent neural dynamics," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pp. 104–111, 2018.

[3] C. Pehlevan and D. Chklovskii, "A normative theory of adaptive dimensionality reduction in neural networks," in *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds.), pp. 2269–2277, Curran Associates, Inc., 2015.