Correlative Information Maximization: A Biologically Plausible Approach to Supervised Deep Neural Networks without Weight Symmetry

Neurips 2023

Bariscan Bozkurt[1,2,3] **Cengiz Pehlevan**[4,5] **Alper T. Erdogan**[1,2]

[1] KUIS AI Center, Koc University, Istanbul, Turkey
[2] Electrical and Electronics Engineering Department, Koc University, Istanbul, Turkey
[3] Gatsby Computational Neuroscience Unit, UCL, United Kingdom
[4] John A. Paulson School of Engineering & Applied Sciences and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA
[5] Kempner Institute for the Study of Natural and Artificial Intelligence

Goal

Introduce a biologically plausible neural network framework grounded on information theory offering,

- ✦ A principled solution to the **weight symmetry** problem,

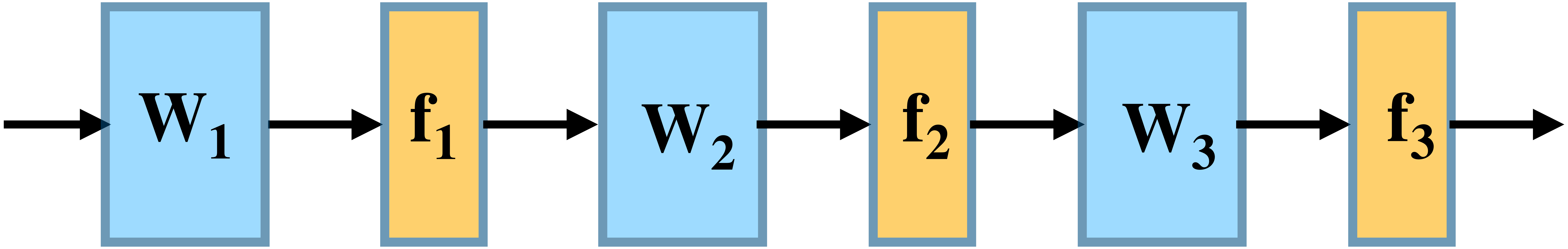
- ✦ A normative approach for deriving networks with **multi-compartment neurons**.

Outline

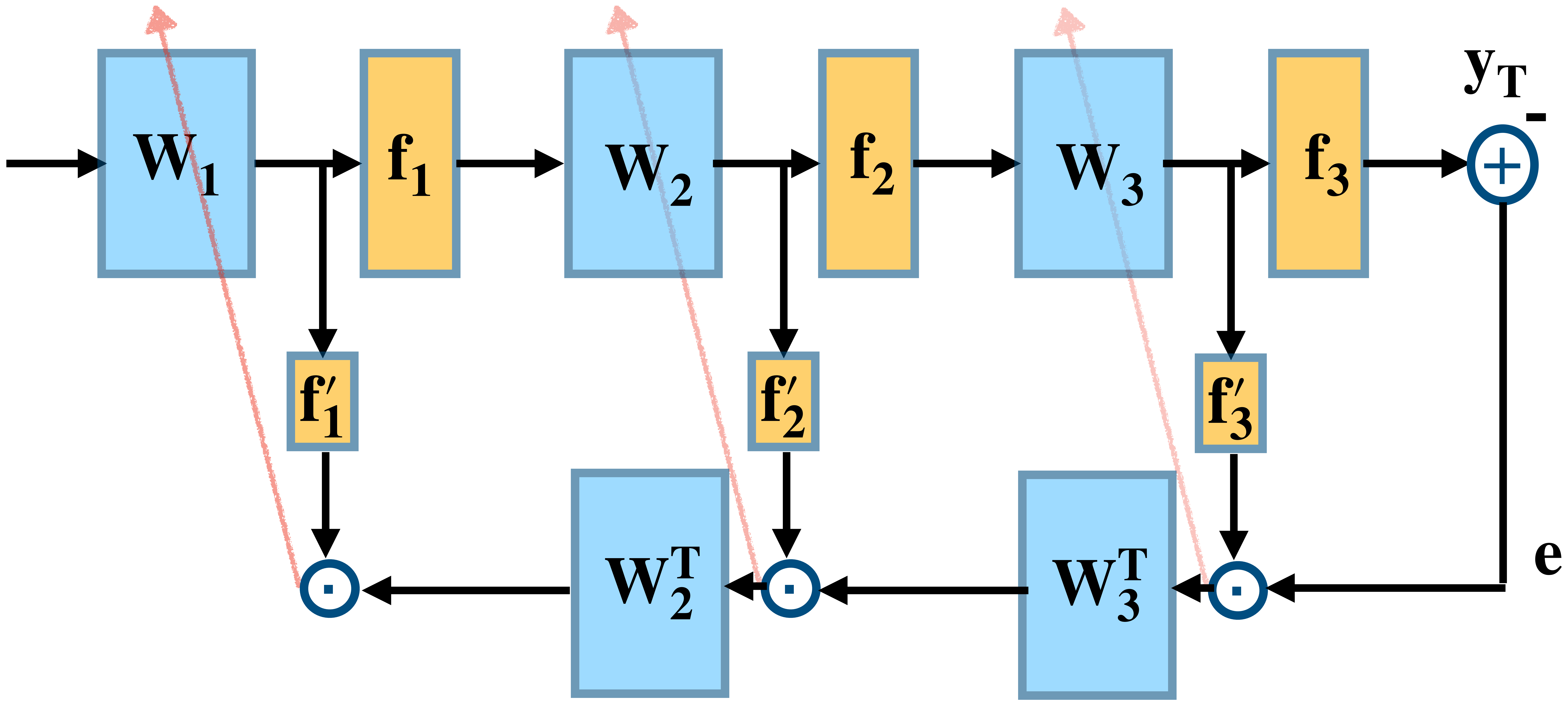
- ✦ Weight Symmetry (Transport) problem
- ✦ **Cor**relative **Info**rmation **Max**imization (CorInfoMax) criterion
- ✦ Network Data Model
- ✦ Network Structure and Dynamics
- ✦ Numerical Examples

Backpropagation: Weight Symmetry Issue

A Multi-Layer Neural Network



Backpropagation: Weight Symmetry Issue



Correlative Information Maximization (CorInfoMax) Criterion

Correlative Mutual Information (CMI)

$\mathbf{a}, \mathbf{b}$: Random vectors

$$\overrightarrow{I}^{(\epsilon_k)}(\mathbf{a}, \mathbf{b}) = \frac{1}{2} \log \det(\hat{\mathbf{R}}_{\mathbf{a}}[t] + \epsilon_k \mathbf{I}) - \frac{1}{2} \log \det(\hat{\mathbf{R}}_{\mathbf{e}_{\mathbf{a}|\mathbf{b}}}[t] + \epsilon_k \mathbf{I})$$

The correlation matrix of $\mathbf{a}$

The error of the best linear MMSE estimator of $\mathbf{a}$ from $\mathbf{b}$

Correlative Information Maximization (CorInfoMax) Criterion

Correlative Mutual Information (CMI)

$\mathbf{a}, \mathbf{b}$: Random vectors

$$I^{(\epsilon_k)}(\mathbf{a}, \mathbf{b}) = \frac{1}{2}\log\det(\hat{\mathbf{R}}_{\mathbf{a}}[t] + \epsilon_k \mathbf{I}) - \frac{1}{2}\log\det(\hat{\mathbf{R}}_{\mathbf{e}_{\mathbf{a}|\mathbf{b}}}[t] + \epsilon_k \mathbf{I})$$

The correlation matrix of $\mathbf{a}$

The error of the best linear MMSE estimator of $\mathbf{a}$ from $\mathbf{b}$

Maximize $I^{(\epsilon_k)}(\mathbf{a}, \mathbf{b})$ $\Longrightarrow$ Maximize correlation (linear dependence) between $\mathbf{a}$ and $\mathbf{b}$

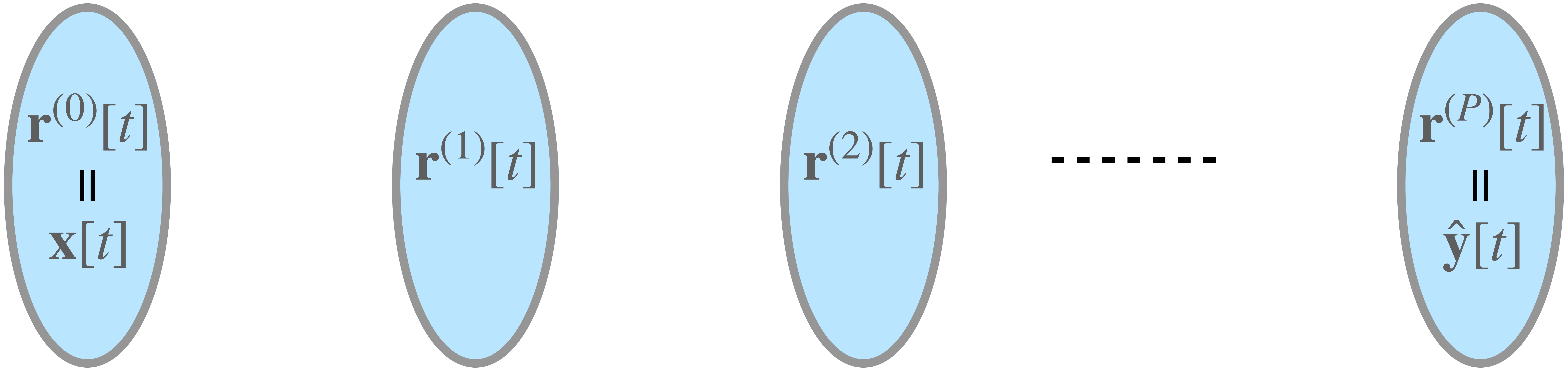
Network Data Model

$\mathbf{x}[t]$: input sample

Network Data Model

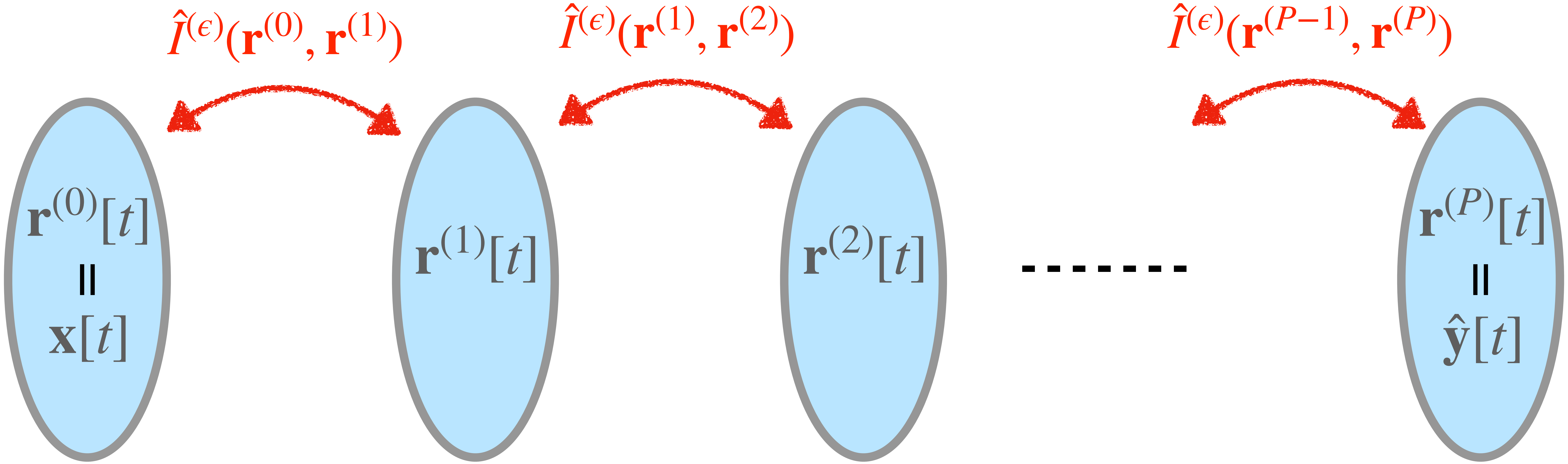
$\mathbf{x}[t]$: input sample

$\mathbf{r}^{(k)}[t]$: layer activations



Correlative Information Maximization Criterion

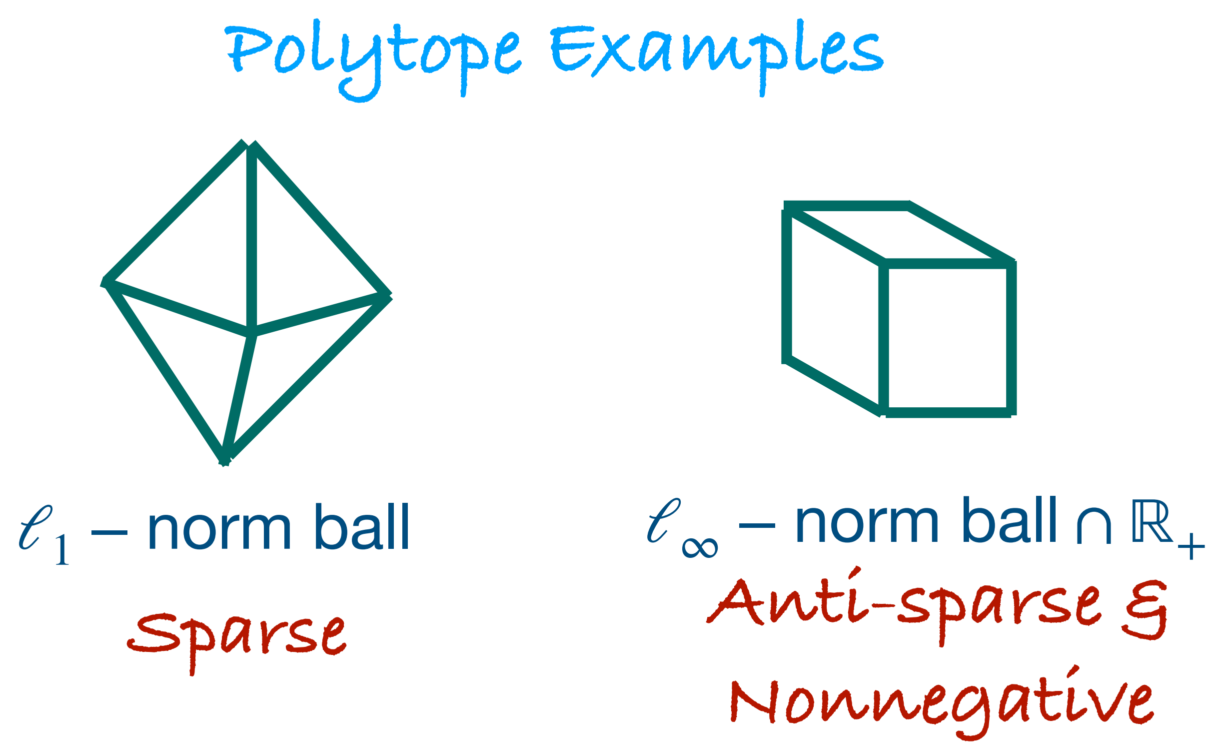
Objective: $\sum_{k=0}^{P-1} \hat{I}^{(\epsilon)}(\mathbf{r}^{(k)}, \mathbf{r}^{(k+1)})$

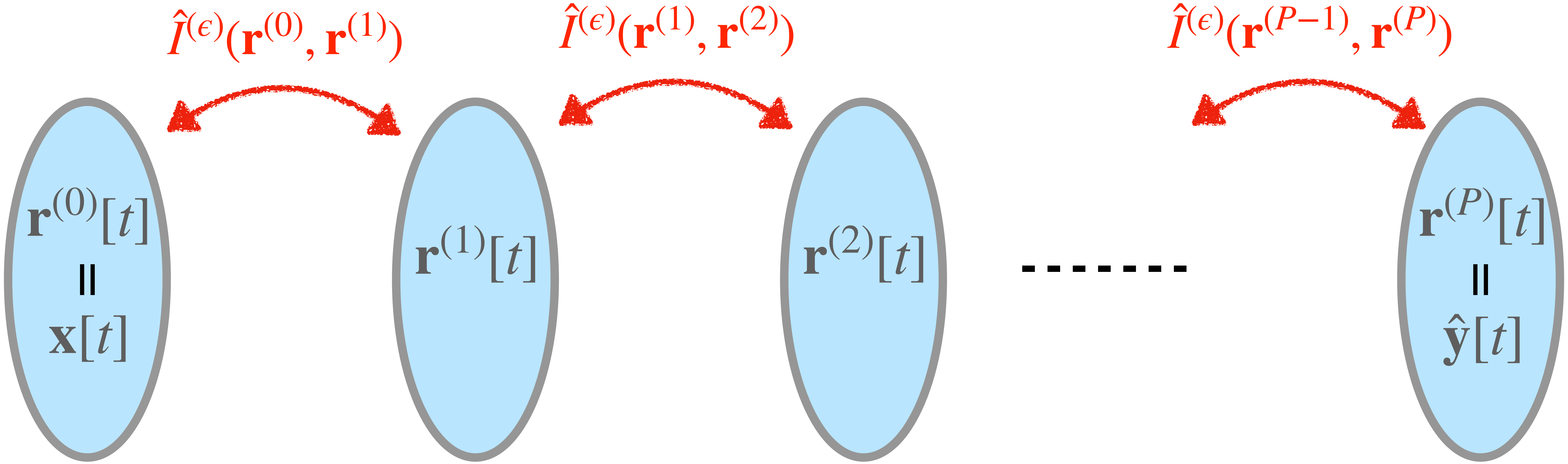


Correlative Information Maximization Criterion

Objective: $\sum_{k=0}^{P-1} \hat{I}^{(\epsilon)}(\mathbf{r}^{(k)}, \mathbf{r}^{(k+1)})$

Constraints: $\mathbf{r}^{(k)} \in \mathscr{P}_k$

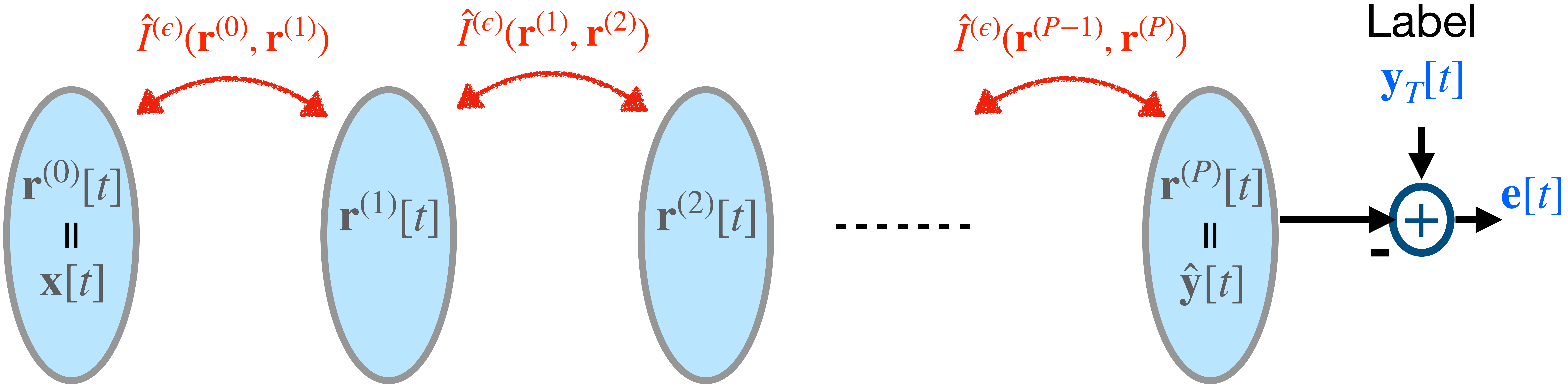




Correlative Information Maximization Criterion

Objective: $$\sum_{k=0}^{P-1} \hat{I}^{(\epsilon)}(\mathbf{r}^{(k)}, \mathbf{r}^{(k+1)})$$

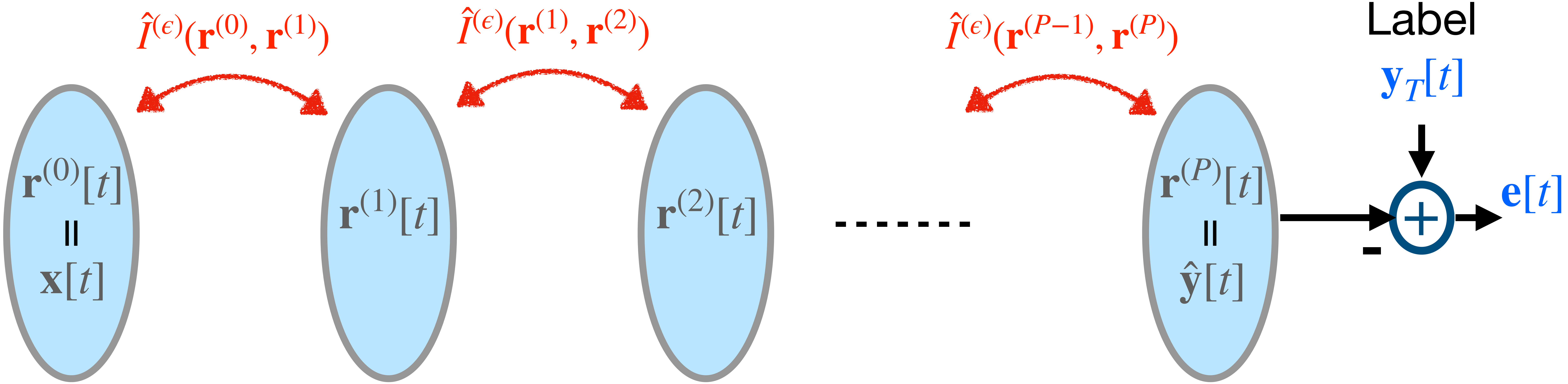
Constraints: $\mathbf{r}^{(k)} \in \mathscr{P}_k$



Correlative Information Maximization Criterion

Objective: $$\sum_{k=0}^{P-1} \hat{I}^{(\epsilon)}(\mathbf{r}^{(k)}, \mathbf{r}^{(k+1)}) - \frac{\beta}{2}\|\mathbf{y}_T[t] - \mathbf{r}^{(P)}[t]\|_2^2$$

Constraints: $$\mathbf{r}^{(k)} \in \mathcal{P}_k$$



Correlative Information Maximization Criterion

Two Alternative Expressions for the Correlative Mutual Information (CMI):

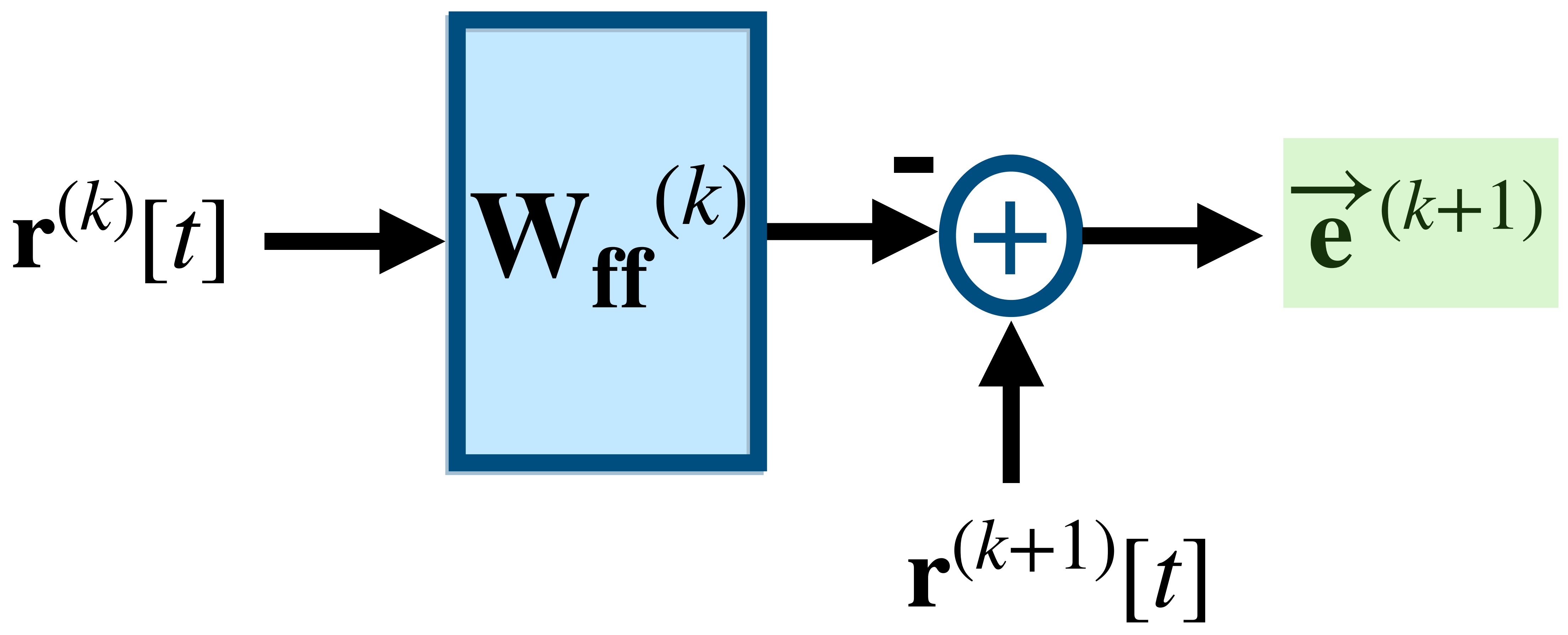
Correlative Information Maximization Criterion

Two Alternative Expressions for the Correlative Mutual Information (CMI):

Alternative 1:

$$\overrightarrow{\hat{I}^{(\epsilon_k)}}(\mathbf{r}^{(k)}, \mathbf{r}^{(k+1)})[t] = \frac{1}{2}\log\det(\hat{\mathbf{R}}_{\mathbf{r}^{(k+1)}}[t] + \epsilon_k \mathbf{I}) - \frac{1}{2}\log\det(\hat{\mathbf{R}}_{\overrightarrow{\mathbf{e}}_*^{(k+1)}}[t] + \epsilon_k \mathbf{I})$$

where



Forward prediction error in predicting $\mathbf{r}^{(k+1)}[t]$ from $\mathbf{r}^{(k)}[t]$

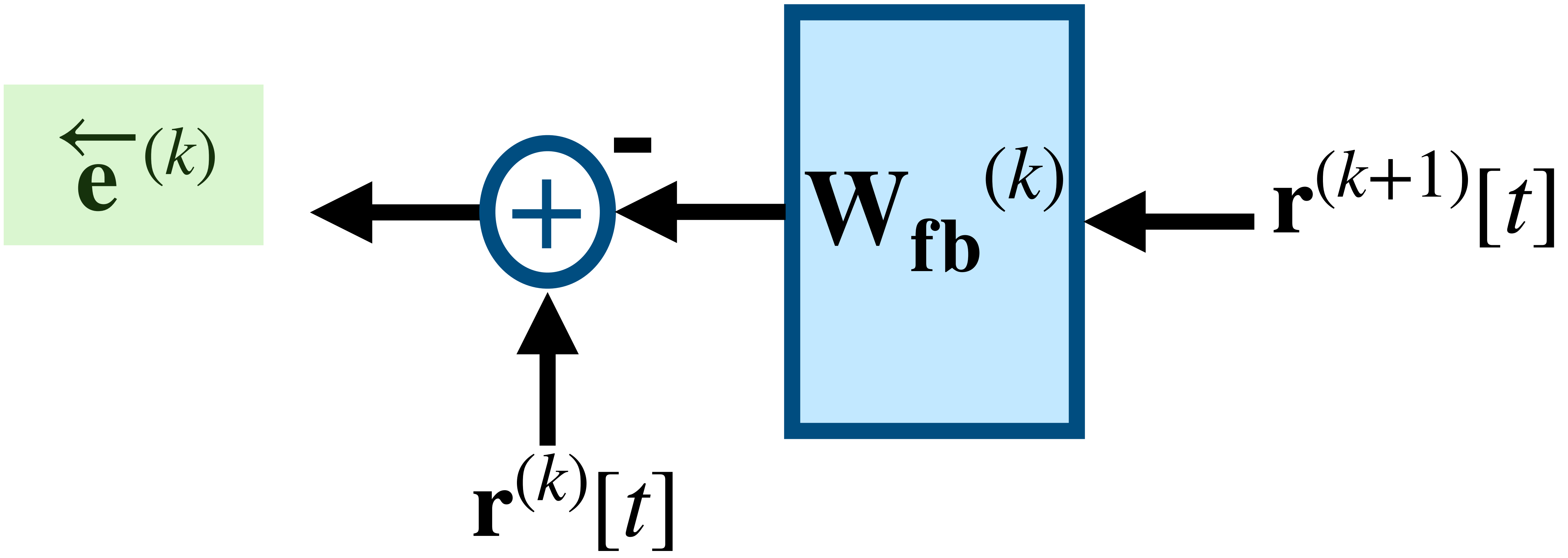
Correlative Information Maximization Criterion

Two Alternative Expressions for the Correlative Mutual Information (CMI):

Alternative 2:

$$\overleftarrow{\hat{I}}^{(\epsilon_k)}(\mathbf{r}^{(k)}, \mathbf{r}^{(k+1)})[t] = \frac{1}{2}\log\det(\hat{\mathbf{R}}_{\mathbf{r}^{(k)}}[t] + \epsilon_k \mathbf{I}) - \frac{1}{2}\log\det(\hat{\mathbf{R}}_{\overleftarrow{\mathbf{e}}_*^{(k)}}[t] + \epsilon_k \mathbf{I})$$

where

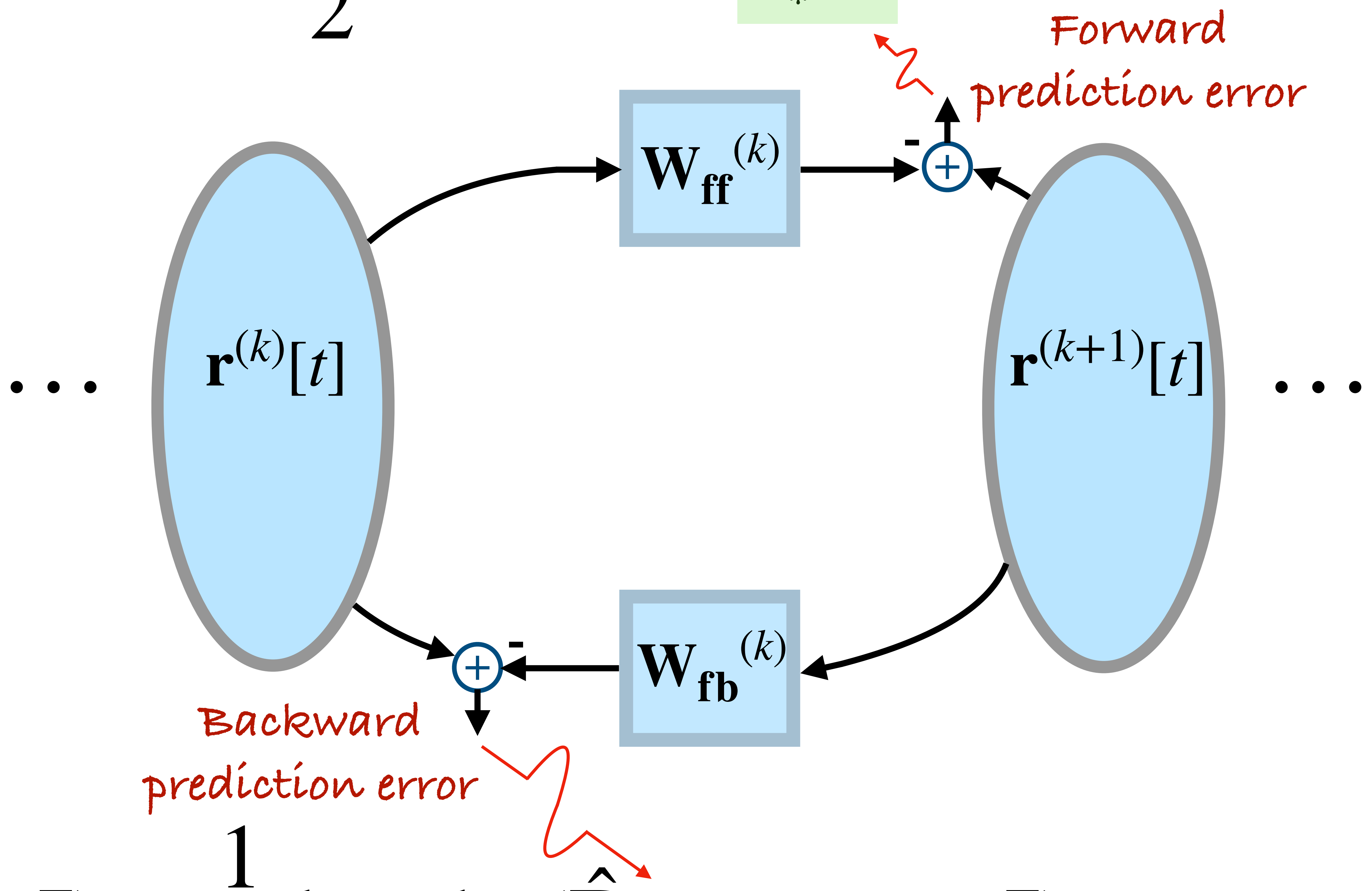


Backward prediction error in predicting $\mathbf{r}^{(k)}[t]$ from $\mathbf{r}^{(k+1)}[t]$

Correlative Information Maximization Criterion

Two Alternative Expressions for the Correlative Mutual Information (CMI):

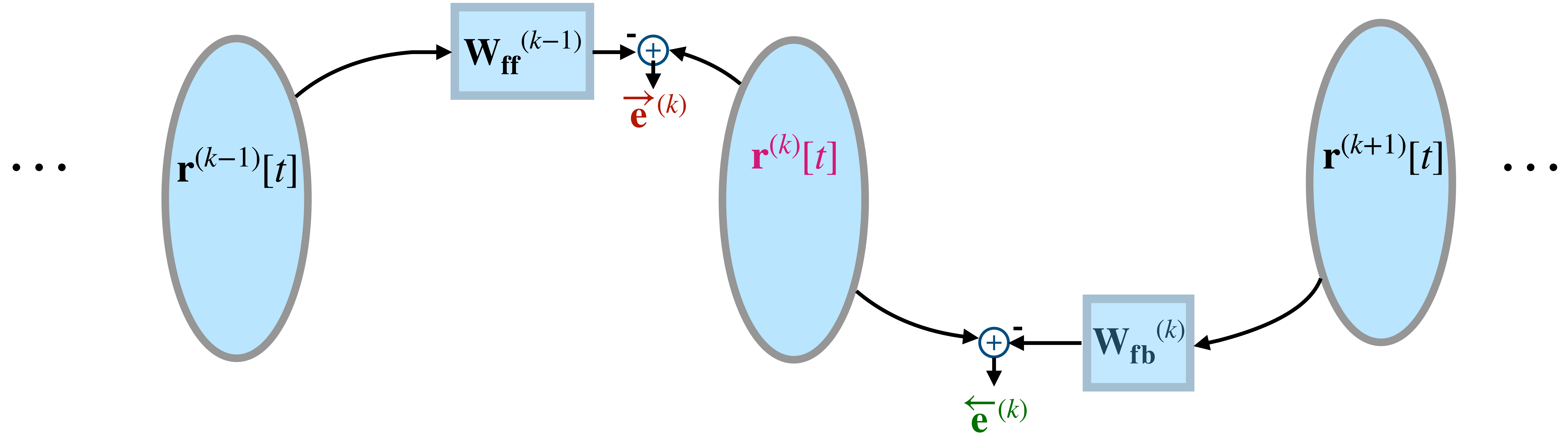
$$\overrightarrow{\hat{I}^{(\epsilon_k)}}(\mathbf{r}^{(k)}, \mathbf{r}^{(k+1)})[t] = \frac{1}{2}\log\det(\hat{\mathbf{R}}_{\mathbf{r}^{(k+1)}}[t] + \epsilon_k \mathbf{I}) - \frac{1}{2}\log\det(\hat{\mathbf{R}}_{\overrightarrow{\mathbf{e}}_*^{(k+1)}}[t] + \epsilon_k \mathbf{I})$$



$$\overleftarrow{\hat{I}^{(\epsilon_k)}}(\mathbf{r}^{(k)}, \mathbf{r}^{(k+1)})[t] = \frac{1}{2}\log\det(\hat{\mathbf{R}}_{\mathbf{r}^{(k)}}[t] + \epsilon_k \mathbf{I}) - \frac{1}{2}\log\det(\hat{\mathbf{R}}_{\overleftarrow{\mathbf{e}}_*^{(k)}}[t] + \epsilon_k \mathbf{I})$$

Correlative Information Maximization Criterion

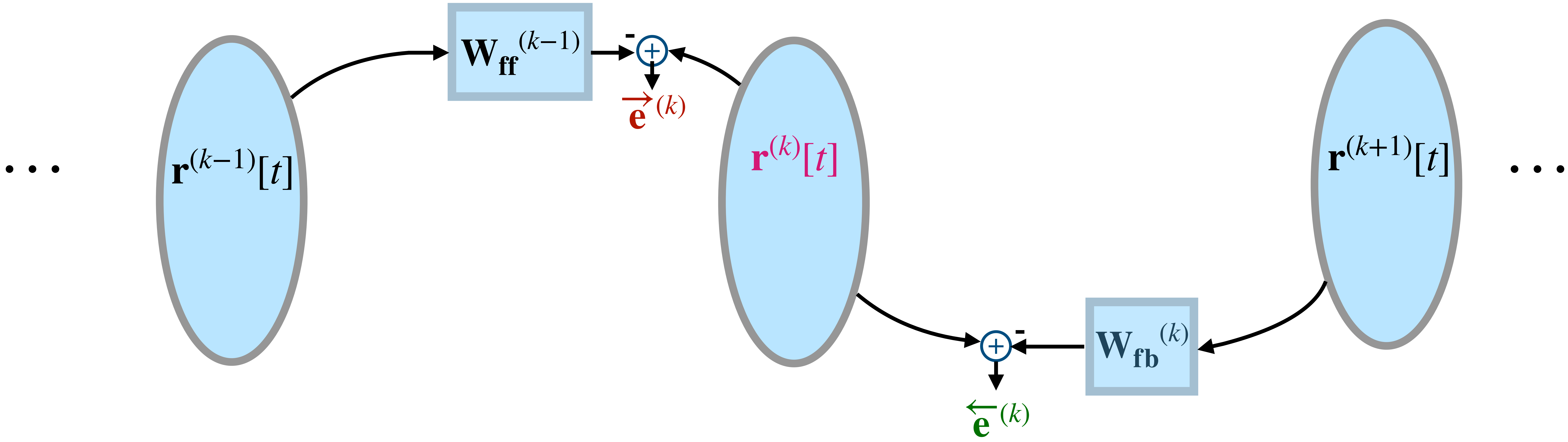
Maximization of CMI: Gradient wrt $\mathbf{r}^{(k)}[t]$



$$\hat{J}_k(\mathbf{r}^{(k)}[t]) = \overrightarrow{\hat{I}^{(\epsilon_{k-1})}}(\mathbf{r}^{(k-1)}, \mathbf{r}^{(k)})[t] + \overleftarrow{\hat{I}^{(\epsilon_k)}}(\mathbf{r}^{(k)}, \mathbf{r}^{(k+1)})[t]$$

Correlative Information Maximization Criterion

Maximization of CMI: Gradient wrt $\mathbf{r}^{(k)}[t]$



$$\nabla_{\mathbf{r}^{(k)}} \hat{J}_k(\mathbf{r}^{(k)}[t]) = \nabla_{\mathbf{r}^{(k)}} \hat{I}^{\overrightarrow{(\epsilon_{k-1})}}(\mathbf{r}^{(k-1)}, \mathbf{r}^{(k)})[t] + \nabla_{\mathbf{r}^{(k)}} \hat{I}^{\overleftarrow{(\epsilon_k)}}(\mathbf{r}^{(k)}, \mathbf{r}^{(k+1)})[t]$$

Network Structure and Neural Dynamics

Multi-compartmental neural network

Projected gradient ascent based neural-dynamics:

Sample index — Continuous time index

$$\tau_{\mathbf{u}} \frac{d\mathbf{u}^{(k)}[t;s]}{ds} = -\, g_{lk}\mathbf{u}^{(k)}[t;s] + g_{A,k}(\mathbf{v}_A^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s]) + g_{B,k}(\mathbf{v}_B^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s])$$

$$\mathbf{r}^{(\mathbf{k})}[t;s] = \sigma_+(\mathbf{u}^{(k)}[t;s])$$

where

Network Structure and Neural Dynamics

Multi-compartmental neural network

Projected gradient ascent based neural-dynamics:

Sample index — Continuous time index

$$\tau_{\mathbf{u}} \frac{d\mathbf{u}^{(k)}[t;s]}{ds} = -\,g_{lk}\mathbf{u}^{(k)}[t;s] + g_{A,k}(\mathbf{v}_A^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s]) + g_{B,k}(\mathbf{v}_B^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s])$$

$$\mathbf{r}^{(\mathbf{k})}[t;s] = \sigma_+(\mathbf{u}^{(k)}[t;s])$$

where

$\mathbf{u}^{(k)}[t,s]$: Somatic compartment membrane potential

Network Structure and Neural Dynamics

Multi-compartmental neural network

Projected gradient ascent based neural-dynamics:

Sample index → $[t; s]$; Continuous time index → s

$$\tau_{\mathbf{u}} \frac{d\mathbf{u}^{(k)}[t;s]}{ds} = -g_{lk}\mathbf{u}^{(k)}[t;s] + g_{A,k}(\mathbf{v}_A^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s]) + g_{B,k}(\mathbf{v}_B^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s])$$

$$\mathbf{r}^{(\mathbf{k})}[t;s] = \sigma_+(\mathbf{u}^{(k)}[t;s])$$

where

$\mathbf{u}^{(k)}[t,s]$: Somatic compartment membrane potential

$\mathbf{v}_A^{(k)}[t,s] = \mathbf{M}^{(k)}[t]\mathbf{r}^{(k)}[t;s] + \mathbf{W}_{fb}^{(k)}[t]\mathbf{r}^{(k+1)}[t;s]$ Apical compartment membrane potential

Network Structure and Neural Dynamics

Multi-compartmental neural network

Projected gradient ascent based neural-dynamics:

Sample index → s ; Continuous time index → t

$$\tau_{\mathbf{u}} \frac{d\mathbf{u}^{(k)}[t;s]}{ds} = -g_{lk}\mathbf{u}^{(k)}[t;s] + g_{A,k}(\mathbf{v}_A^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s]) + g_{B,k}(\mathbf{v}_B^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s])$$

$$\mathbf{r}^{(\mathbf{k})}[t;s] = \sigma_+(\mathbf{u}^{(k)}[t;s])$$

where

$\mathbf{u}^{(k)}[t,s]$: Somatic compartment membrane potential

$\mathbf{v}_B^{(k)}[t,s] = \mathbf{W}_{ff}^{(k-1)}[t]\mathbf{r}^{(k-1)}[t;s]$ Basal compartment membrane potential

$\mathbf{v}_A^{(k)}[t,s] = \mathbf{M}^{(k)}[t]\mathbf{r}^{(k)}[t;s] + \mathbf{W}_{fb}^{(k)}[t]\mathbf{r}^{(k+1)}[t;s]$ Apical compartment membrane potential

Network Structure and Neural Dynamics

Multi-compartmental neural network

Projected gradient ascent based neural-dynamics:

Sample index — Continuous time index

$$\tau_{\mathbf{u}} \frac{d\mathbf{u}^{(k)}[t;s]}{ds} = -g_{lk}\mathbf{u}^{(k)}[t;s] + g_{A,k}(\mathbf{v}_A^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s]) + g_{B,k}(\mathbf{v}_B^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s])$$

$$\mathbf{r}^{(\mathbf{k})}[t;s] = \sigma_+(\mathbf{u}^{(k)}[t;s])$$

where

$g_{A,k} = \frac{1}{\epsilon_{k-1}}$ Apical-Soma Conductance

$g_{B,k} = \frac{1}{\epsilon_k}$ Basal-Soma Conductance

g_{lk} : Leakage Conductance

Network Structure and Neural Dynamics

Multi-compartmental neural network

Projected gradient ascent based neural-dynamics:

Sample index → t; Continuous time index → s

$$\tau_{\mathbf{u}}\frac{d\mathbf{u}^{(k)}[t;s]}{ds} = -g_{lk}\mathbf{u}^{(k)}[t;s] + g_{A,k}(\mathbf{v}_A^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s]) + g_{B,k}(\mathbf{v}_B^{(k)}[t;s] - \mathbf{u}^{(k)}[t;s])$$

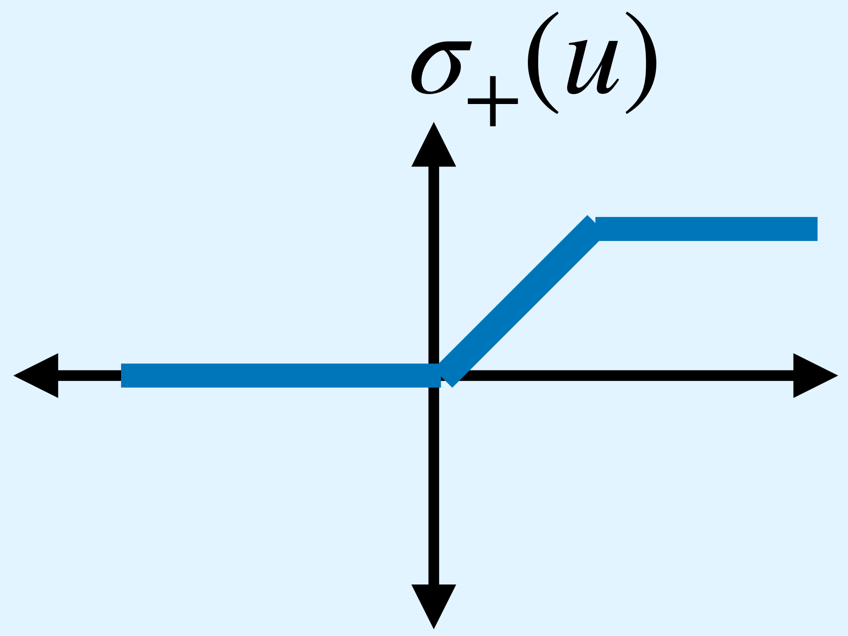
$$\mathbf{r}^{(\mathbf{k})}[t;s] = \sigma_+(\mathbf{u}^{(k)}[t;s])$$

where

$g_{A,k} = \frac{1}{\epsilon_{k-1}}$ Apical-Soma Conductance

$g_{B,k} = \frac{1}{\epsilon_k}$ Basal-Soma Conductance

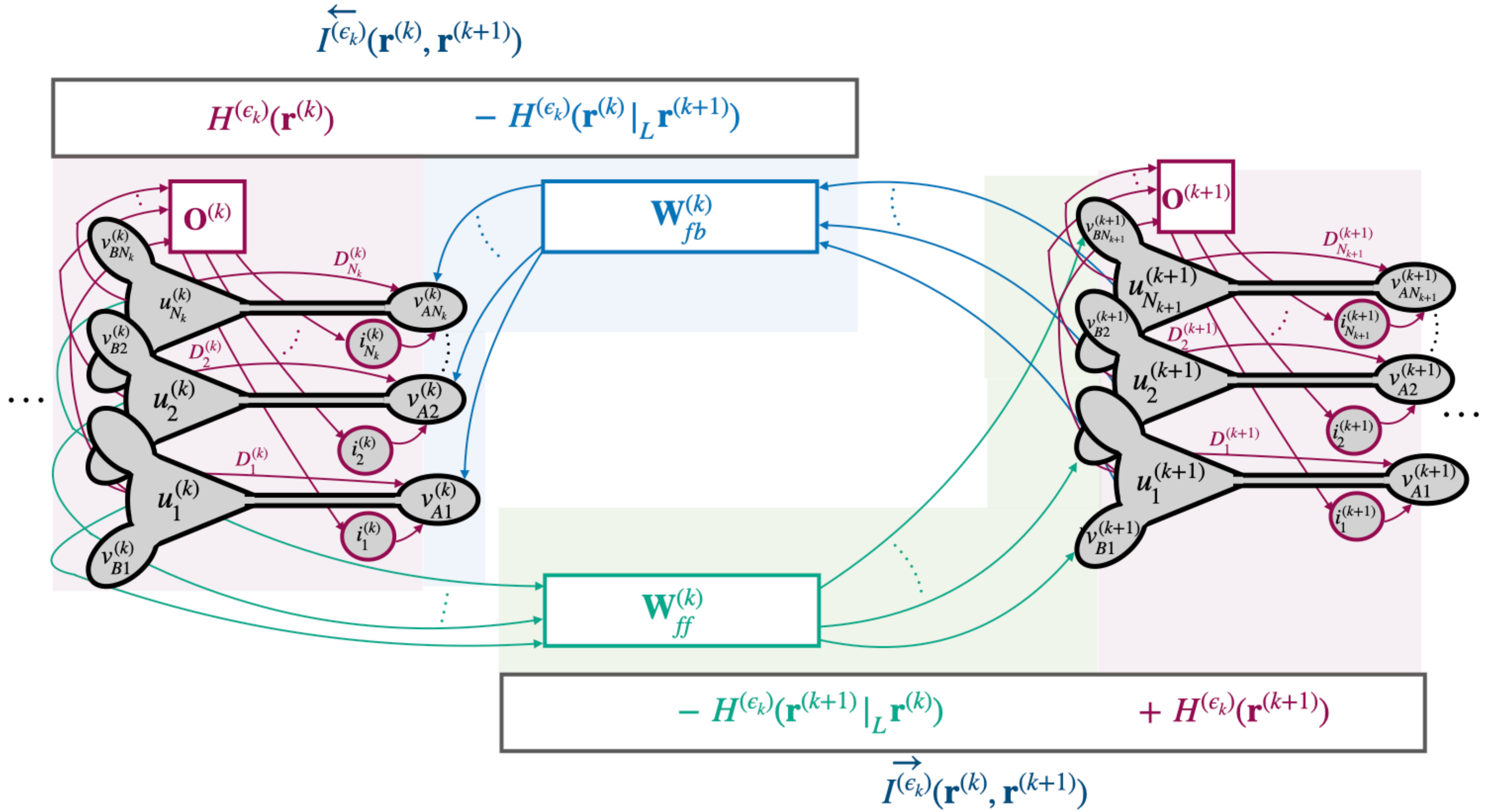
g_{lk} : Leakage Conductance



Projection onto Hypercube

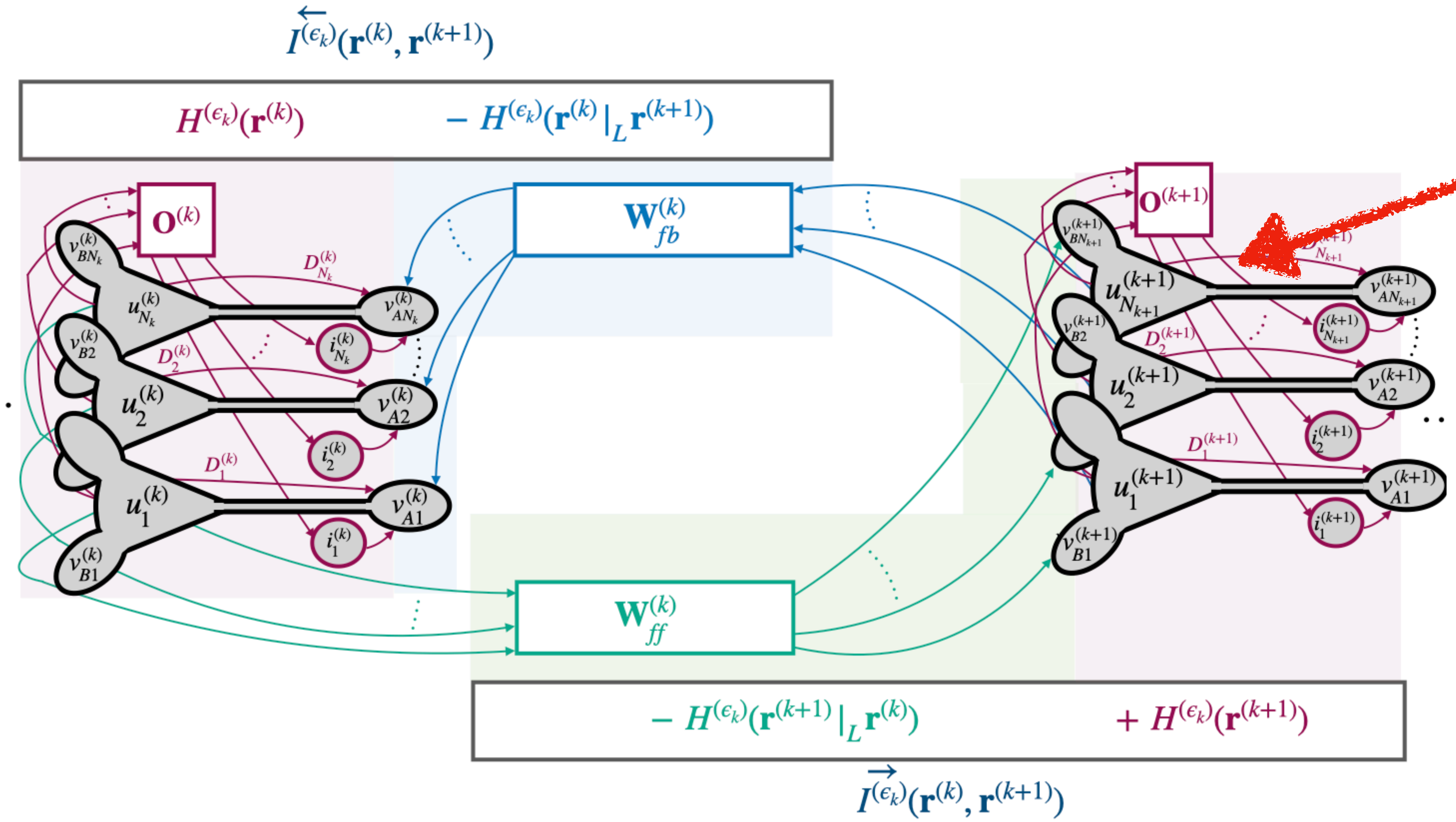
Network Structure and Neural Dynamics

Multi-compartmental neural network



Network Structure and Neural Dynamics

Multi-compartmental neural network

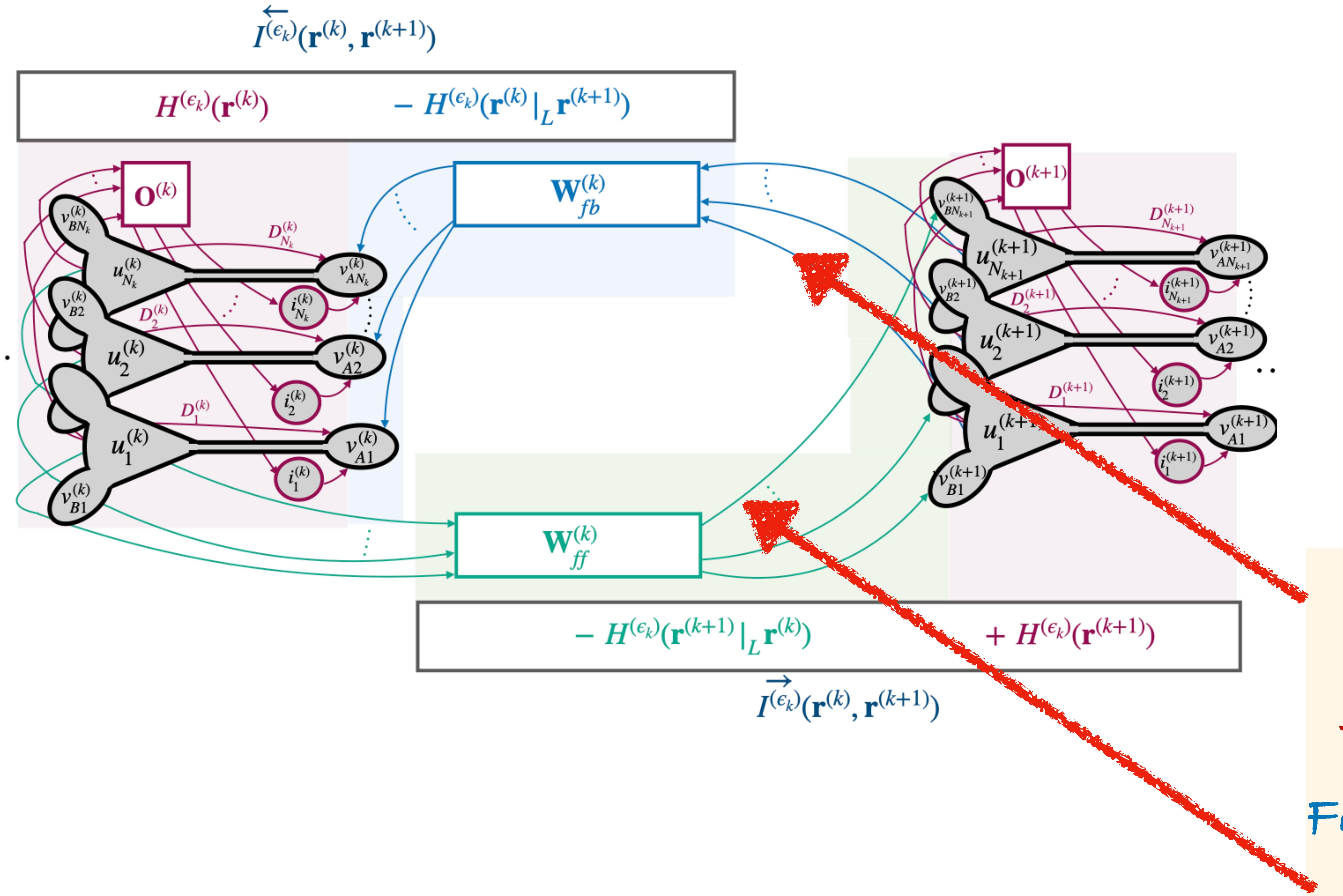


Lateral connections
to maximize layer entropy:

utilization of representation
dimensions, avoid degeneracy

Network Structure and Neural Dynamics

Multi-compartmental neural network



Feedforward/Feedback connections
to minimize conditional entropy:
Facilitate bidirectional information flow,reduce redundancy

Network Structure and Neural Dynamics

Learning Dynamics

$$\delta \mathbf{W}_{ff}^{(k)}[t] \propto \frac{1}{\beta'} \left(\left(\overrightarrow{\mathbf{e}}^{(k+1)}[t] \mathbf{r}^{(k)}[t]^T \right) \Bigg|_{\beta=\beta'} - \left(\overrightarrow{\mathbf{e}}^{(k+1)}[t] \mathbf{r}^{(k)}[t]^T \right) \Bigg|_{\beta=0} \right)$$

$$\delta \mathbf{W}_{fb}^{(k)}[t] \propto \frac{1}{\beta'} \left(\left(\overleftarrow{\mathbf{e}}^{(k)}[t] \mathbf{r}^{(k+1)}[t]^T \right) \Bigg|_{\beta=\beta'} - \left(\overleftarrow{\mathbf{e}}^{(k)}[t] \mathbf{r}^{(k+1)}[t]^T \right) \Bigg|_{\beta=0} \right)$$

$$\mathbf{B}^{(k)}[t+1] = \lambda_{\mathbf{r}}^{-1} (\mathbf{B}^{(k)}[t] - \gamma \mathbf{z}^{(k)}[t] \mathbf{z}^{(k)}[t]^T), \text{ where } \mathbf{z}^{(k)} = \mathbf{B}^{(k)}[t] \mathbf{r}^{(k)}[t] \Big|_{\beta=\beta'}$$

Numerical Examples

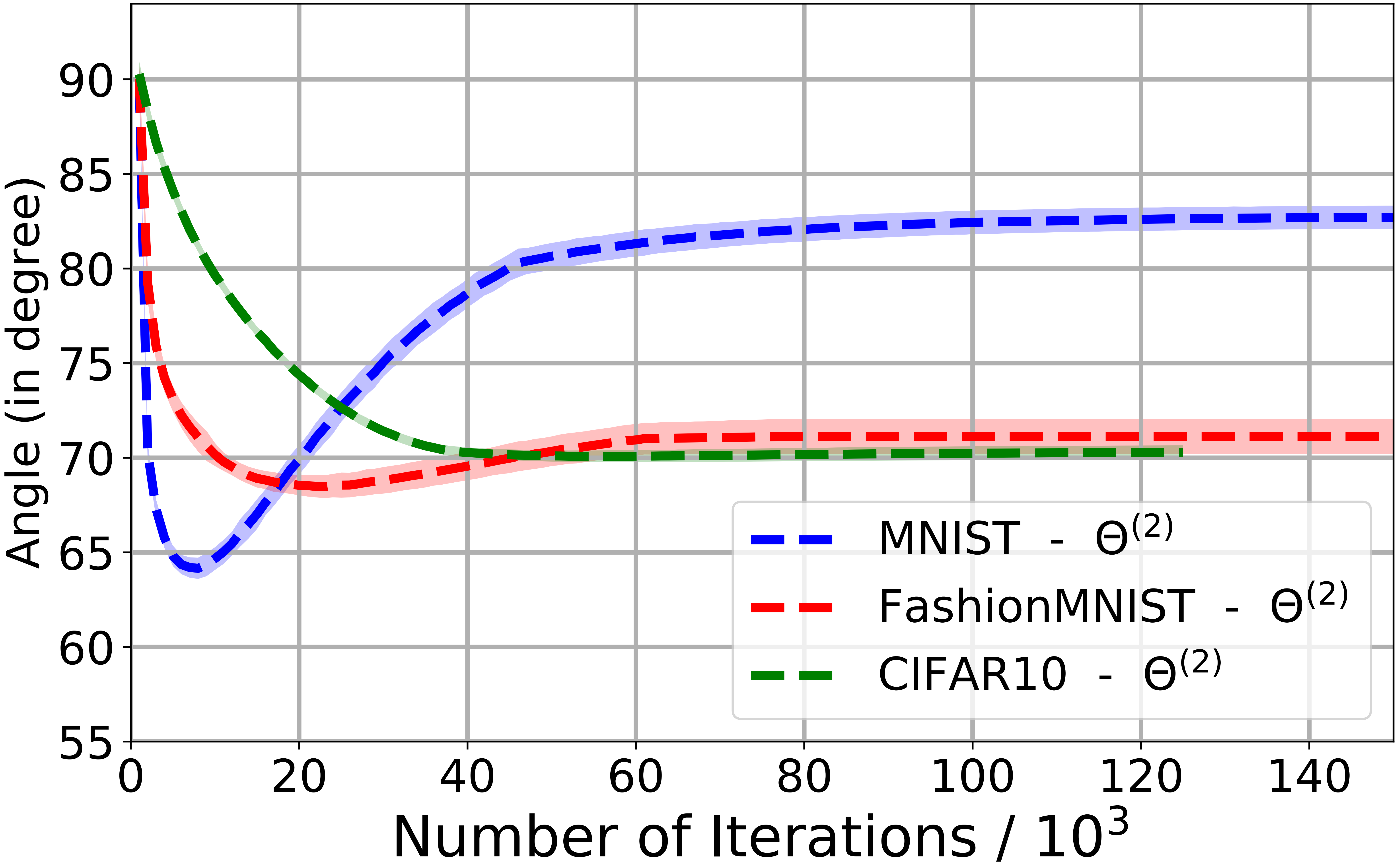
- **MNIST** and **Fashion MNIST** Datasets: Layer sizes of 784, 500,10
- **CIFAR 10** Dataset: Layer sizes of 3072,1000,10

	MNIST	FashionMNIST	CIFAR10
CorInfoMax-$\mathcal{B}_{\infty,+}$	97.62 ± 0.1	88.14 ± 0.3	51.86 ± 0.3
CorInfoMax-$\mathcal{B}_{1,+}$	97.71 ± 0.1	88.09 ± 0.1	51.19 ± 0.4
EP	97.61 ± 0.1	88.06 ± 0.7	49.28 ± 0.5
CSM	98.08 ± 0.1	88.73 ± 0.2	40.79^*
PC	98.17 ± 0.2	89.31 ± 0.4	-
PC-Nudge	97.71 ± 0.1	88.49 ± 0.3	48.58 ± 0.7
Feedback Alignment (with MSE Loss)	97.99 ± 0.03	88.72 ± 0.5	50.75 ± 0.4
Feedback Alignment (with CrossEntropy Loss)	97.95 ± 0.08	88.38 ± 0.9	52.37 ± 0.4
BP (with MSE Loss)	97.58 ± 0.01	88.39 ± 0.1	52.75 ± 0.1
BP (with CrossEntropy Loss)	98.27 ± 0.03	89.41 ± 0.2	53.96 ± 0.3

Numerical Examples

Confirming Asymmetry: Angle between feedforward and the transpose of the feedback weights

$$\Theta^{(k)} = \arccos\left(\frac{\mathrm{Tr}\left(\mathbf{W}_{ff}^{(k)}\mathbf{W}_{fb}^{(k)}\right)}{\|\mathbf{W}_{ff}^{(k)}\|_F\|\mathbf{W}_{fb}^{(k)}\|_F}\right)$$



Conclusions

CorInfoMax is offered as an information theory-based principled framework:

- Networks of segregated neurons with recurrent and asymmetric feedback/ feedforward connections governed by local learning rules naturally emerge,

- Useful in obtaining potential insights such as

  - the role of lateral connections in embedding space expansion and avoiding degeneracy,

  - feedback and feedforward connections for prediction to reduce redundancy,

  - activation functions/interneurons to shape feature space and compress.

Thank You!