Spontaneous symmetry breaking in machine learning

Haiping Huang

RIKEN, Brain Science Institute, Japan

Machine Learning and Many-body Physics, KITS, Beijing
arXiv:1608.03714, arXiv:1612.01717, and arXiv:1703.07943

July 4, 2017

Outline

1. Introduction
2. A simple model
   - Bethe approximation
   - Replica theory
3. Phase transitions
4. How cold is a dataset
5. Roles of zero synapses
6. Summary

Outline

1. Introduction
2. A simple model
   - Bethe approximation
   - Replica theory
3. Phase transitions
4. How cold is a dataset
5. Roles of zero synapses
6. Summary

Outline

1. Introduction
2. A simple model
   - Bethe approximation
   - Replica theory
3. Phase transitions
4. How cold is a dataset
5. Roles of zero synapses
6. Summary

Outline

1. Introduction
2. A simple model
   - Bethe approximation
   - Replica theory
3. Phase transitions
4. How cold is a dataset
5. Roles of zero synapses
6. Summary

Outline

1. Introduction
2. A simple model
   - Bethe approximation
   - Replica theory
3. Phase transitions
4. How cold is a dataset
5. Roles of zero synapses
6. Summary

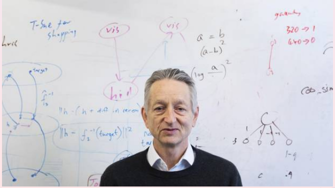
Outline

1. Introduction
2. A simple model
   - Bethe approximation
   - Replica theory
3. Phase transitions
4. How cold is a dataset
5. Roles of zero synapses
6. Summary

Unsupervised learning

Almost every deep-learning product in commercial use today uses supervised learning, meaning that the neural net is trained with labeled data (like the images assembled by ImageNet). With unsupervised learning, by contrast, a neural net is shown unlabeled data and asked simply to look for recurring patterns. Researchers would love to master unsupervised learning one day because then machines could teach themselves about the world from vast stores of data that are unusable today—making sense of the world almost totally on their own, like infants.

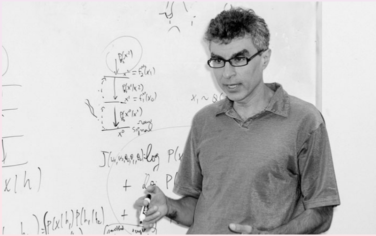
—Geoffrey E. Hinton



Unsupervised learning

If you think about it, scientists are doing unsupervised learning: observing the world, coming up with explanatory models, testing them by collecting more (targeted, though) observations, and continuously trying to improve our causal model of how the world around us works.

—Yoshua Bengio



Key motivations

However, to understand deep learning as a whole is extremely difficult and highly challenging.

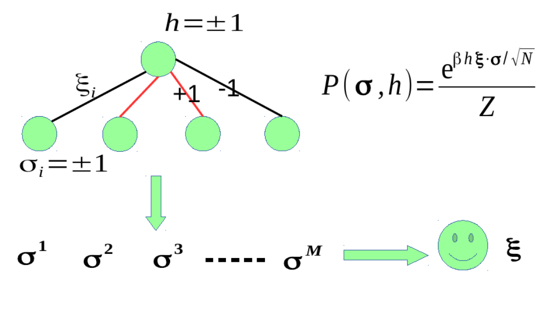
- How does learning improve with data size?
- How many data are required to learn a feature?
- What key factors determine the success of unsupervised learning?

Simple but non-trivial

JW Gibbs (1881)

One of the principal objects of theoretical research...is to find the point of view from which the subject appears in its greatest simplicity.

One-bit Restricted Boltzmann Machine

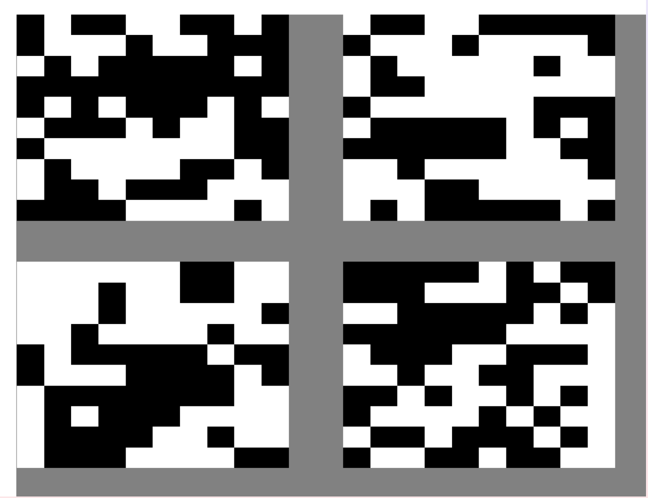


$$P(\boldsymbol{\sigma},h)=\frac{e^{\beta h \boldsymbol{\xi}\cdot\boldsymbol{\sigma}/\sqrt{N}}}{Z}$$

The posterior distribution of the feature vector:

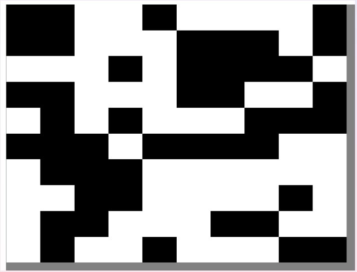
$$P(\boldsymbol{\xi}|\{\boldsymbol{\sigma}^a\}) \propto \prod_a P(\boldsymbol{\sigma}^a|\boldsymbol{\xi}) = \frac{1}{Z}\prod_a \cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}^a\right), \quad (1)$$

An intuitive example: data samples (10 by 10)

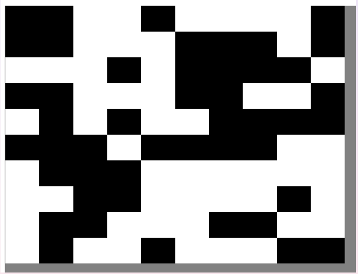


An intuitive example: after showing $10N$ samples

true



guess

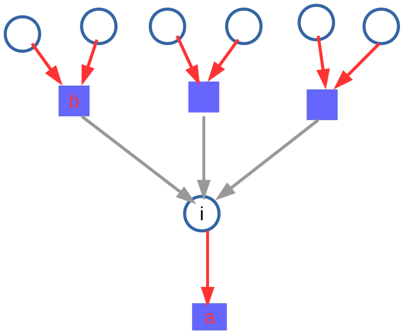


only one bit is different!!

Outline

1. Introduction
2. A simple model
   - Bethe approximation
   - Replica theory
3. Phase transitions
4. How cold is a dataset
5. Roles of zero synapses
6. Summary

Factor graph: Bethe approximation



$$P_{i\to a}(\xi_i) \propto \prod_{b\in\partial i\backslash a} \mu_{b\to i}(\xi_i), \tag{2a}$$

$$\mu_{b\to i}(\xi_i) = \sum_{\{\xi_j|j\in\partial b\backslash i\}} \cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}^{b}\right) \prod_{j\in\partial b\backslash i} P_{j\to b}(\xi_j), \tag{2b}$$

Central limit theorem

$$\mu_{b\to i}(\xi_i) = \sum_{\{\sigma_j | j\in\partial b\backslash i\}} \cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}^b\right) \prod_{j\in\partial b\backslash i} P_{j\to b}(\xi_j) \tag{3}$$

$G_{b\to i} = \frac{1}{\sqrt{N}}\sum_{j\in\partial b\backslash i}\sigma_j^b m_{j\to b}$, $\Xi^2_{b\to i} \simeq \frac{1}{N}\sum_{j\in\partial b\backslash i}(1 - m^2_{j\to b})$.

Huang & Toyoizumi., Phys Rev E (2015).

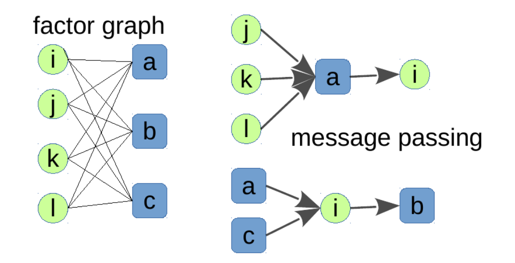
Distributed message passing

Simplified Belief Propagation (sBP):

$$m_{i\to b} = \tanh\left(\sum_{a\in\partial i\backslash b} u_{a\to i}\right), \tag{4a}$$

$$u_{a\to i} = \tanh^{-1}\left(\tanh(\beta G_{a\to i})\tanh(\beta\sigma_i^a/\sqrt{N})\right), \tag{4b}$$

One iteration requires $O(MN)$ computations!



Mean field estimator and entropy

The maximizer of the posterior marginals (MPM) estimator

$$\hat{\xi}_i = \arg\max_{\xi_i} P_i(\xi_i) \tag{5}$$

maximizing the overlap $q = \frac{1}{N}\sum_i \xi_i^{\text{true}}\hat{\xi}_i$.

entropy

$s = \frac{1}{N}\ln\Omega = -\frac{1}{N}\sum_{\xi} P(\xi)\ln P(\xi)$.

—the number of feature vectors consistent with the presented data.

Mean field estimator and entropy

The maximizer of the posterior marginals (MPM) estimator

$$\hat{\xi}_i = \arg\max_{\xi_i} P_i(\xi_i) \tag{5}$$

maximizing the overlap $q = \frac{1}{N}\sum_i \xi_i^{\rm true}\hat{\xi}_i$.

entropy

$s \equiv \frac{1}{N}\ln\Omega = -\frac{1}{N}\sum_{\boldsymbol{\xi}} P(\boldsymbol{\xi})\ln P(\boldsymbol{\xi})$.

—the number of feature vectors consistent with the presented data.

Outline

1. Introduction
2. A simple model
   - Bethe approximation
   - Replica theory
3. Phase transitions
4. How cold is a dataset
5. Roles of zero synapses
6. Summary

Replica theory

Typical behavior: averaged over random features ($\boldsymbol{\xi}^{\rm true}$) and corresponding data ($\{\boldsymbol{\sigma}^a\}$).

$$-\beta f = \lim_{n\to 0, N\to\infty} \frac{\ln \langle Z^n\rangle}{nN}, \tag{6}$$

$$\langle Z^n\rangle = \frac{1}{2^N}\sum_{\{\boldsymbol{\sigma}^a, \boldsymbol{\xi}^{\rm true}\}} P(\{\boldsymbol{\sigma}^a\}|\boldsymbol{\xi}^{\rm true}) \sum_{\{\boldsymbol{\xi}^\gamma\}}\prod_{a,\gamma}\cosh\left(\frac{\beta\boldsymbol{\xi}^\gamma\boldsymbol{\sigma}^a}{\sqrt{N}}\right). \tag{7}$$

Free energy under permutation symmetry of replica matrix:

$$\begin{aligned}-\beta f_{\rm RS} = -q\hat{q} + \frac{\hat{r}(r-1)}{2} + \frac{\alpha\beta^2}{2}(1-r) + \int Dz \ln 2\cosh(\hat{q} + \sqrt{\hat{r}}z) \\ +\alpha e^{-\beta^2/2}\int Dy \int Dt \cosh\beta t \ln\cosh\beta(qt + \sqrt{r-q^2}y).\end{aligned} \tag{8}$$

Huang, JSTAT (2017)

Replica theory

Typical behavior: averaged over random features ($\boldsymbol{\xi}^{\text{true}}$) and corresponding data ($\{\boldsymbol{\sigma}^a\}$).

$$-\beta f = \lim_{n\to 0, N\to\infty} \frac{\ln\langle Z^n\rangle}{nN}, \tag{6}$$

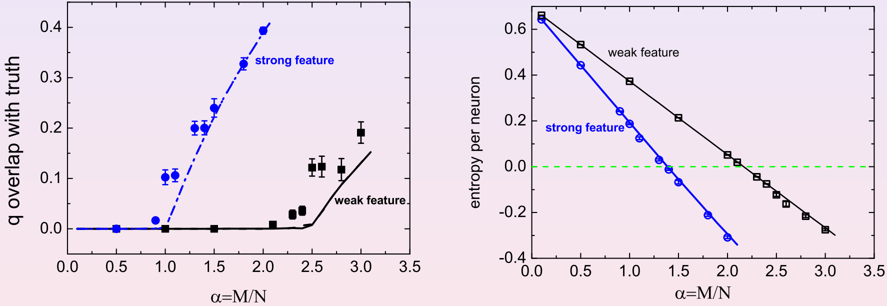
$$\langle Z^n\rangle = \frac{1}{2^N}\sum_{\{\boldsymbol{\sigma}^a,\boldsymbol{\xi}^{\text{true}}\}} P(\{\boldsymbol{\sigma}^a\}|\boldsymbol{\xi}^{\text{true}}) \sum_{\{\boldsymbol{\xi}^\gamma\}}\prod_{a,\gamma}\cosh\left(\frac{\beta\boldsymbol{\xi}^\gamma\boldsymbol{\sigma}^a}{\sqrt{N}}\right). \tag{7}$$

Free energy under permutation symmetry of replica matrix:

$$\begin{aligned}-\beta f_{\text{RS}} = &-q\hat{q} + \frac{\hat{r}(r-1)}{2} + \frac{\alpha\beta^2}{2}(1-r) + \int Dz \ln 2\cosh(\hat{q} + \sqrt{\hat{r}}z) \\ &+\alpha e^{-\beta^2/2}\int Dy\int Dt \cosh\beta t \ln\cosh\beta(qt + \sqrt{r-q^2}y).\end{aligned} \tag{8}$$

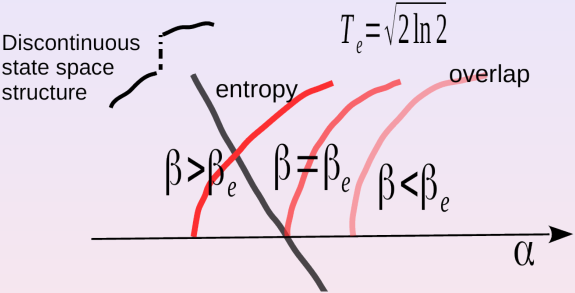
Huang, JSTAT (2017)

Entropy/overlap vs. data size



continuous phase transition $\alpha_c = \beta^{-4} \neq \alpha_{\mathrm{ZE}}$ in general!

Phase diagram

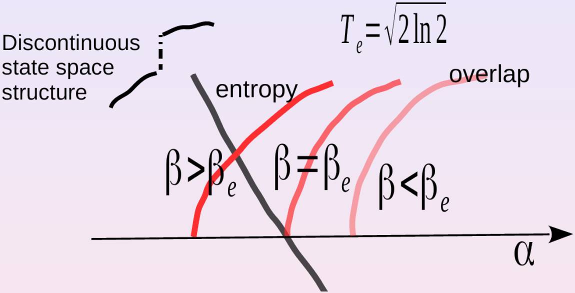


sBP is stable, related to Nishimori condition:

ξ^{true} follows the posterior as well!!

Nishimori, Journal of Physics C: Solid State Physics, 1980; W. Kauzmann, Chem. Rev. (1948).

Phase diagram

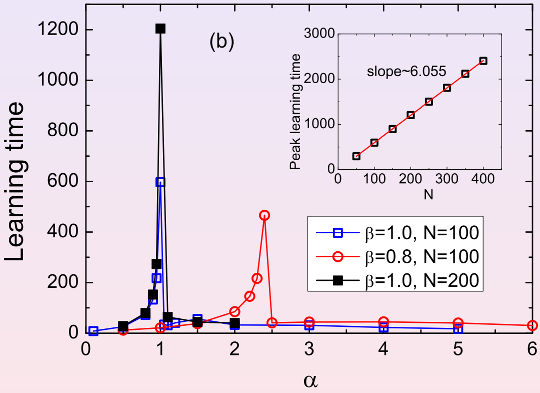


sBP is stable, related to Nishimori condition:

ξ^{true} follows the posterior as well!!

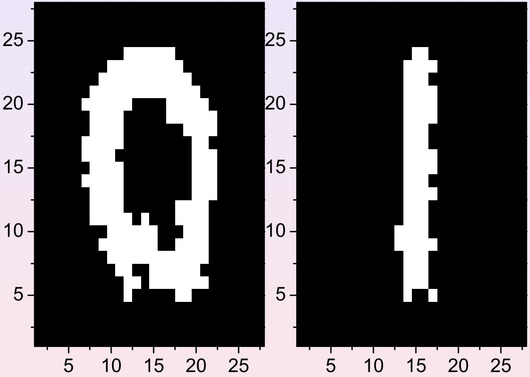
Nishimori, Journal of Physics C: Solid State Physics, 1980; W. Kauzmann, Chem. Rev. (1948).

Easy-hard-easy learning pattern



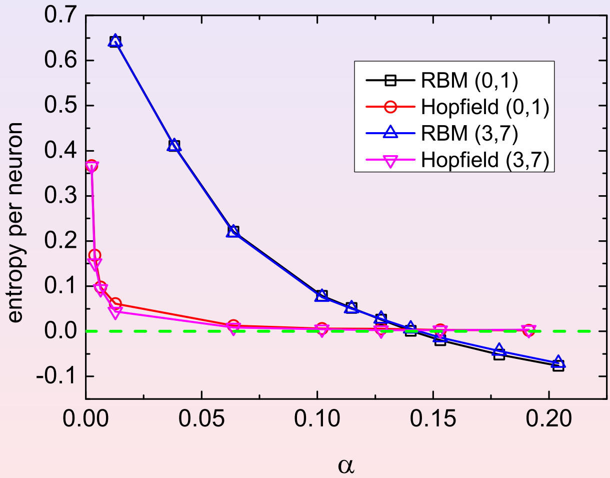
!!!difficult as a linear peak learning time!!!.

Modeling handwriten digits



$N = 28 \times 28, \quad \beta = 1.$

Entropy


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
entropy per neuron
0.00
0.05
0.10
0.15
0.20
α
RBM (0,1)
Hopfield (0,1)
RBM (3,7)
Hopfield (3,7)

How cold is a dataset?

Can we predict feature strength directly from the data?

Temperature of a dataset

The posterior probability of β given the data $\{\boldsymbol{\sigma}^a\}_{a=1}^M$ is given by

$$
\begin{aligned}
P(\beta|\{\boldsymbol{\sigma}^a\}) &= \sum_{\boldsymbol{\xi}} P(\beta,\boldsymbol{\xi}|\{\boldsymbol{\sigma}^a\}) = \frac{\sum_{\boldsymbol{\xi}} P(\{\boldsymbol{\sigma}^a\}|\boldsymbol{\xi},\beta)P_0(\boldsymbol{\xi},\beta)}{\int d\beta \sum_{\boldsymbol{\xi}} P(\{\boldsymbol{\sigma}^a\}|\boldsymbol{\xi},\beta)P_0(\boldsymbol{\xi},\beta)} \\
&= \frac{1}{Z(\{\boldsymbol{\sigma}^a\})} \sum_{\boldsymbol{\xi}} e^{-NM\ln\left(2\cosh(\beta/\sqrt{N})\right)} \prod_a \cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}^a\right) \\
&\propto e^{-M\frac{\beta^2}{2}} Z(\beta,\{\boldsymbol{\sigma}^a\}),
\end{aligned}
\tag{9}
$$

Iterative equation for temperature prediction:

$$
\frac{\partial \ln Z(\beta,\{\boldsymbol{\sigma}^a\})}{\partial\beta} = N\alpha\beta. \tag{10}
$$

Nishimori condition (physics) vs. EM algorithm (statistics)!

Temperature of a dataset

The posterior probability of β given the data $\{\boldsymbol{\sigma}^a\}_{a=1}^M$ is given by

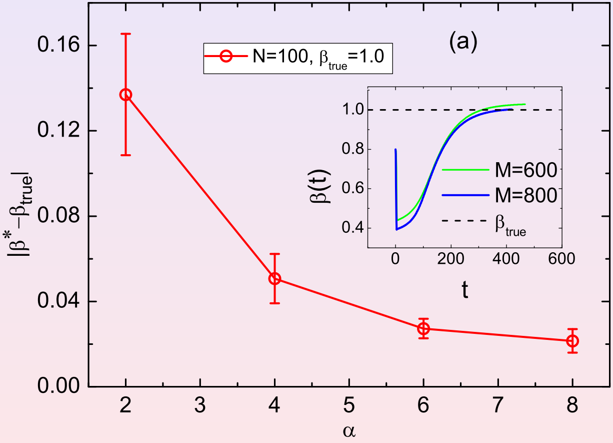
$$
\begin{aligned}
P(\beta|\{\boldsymbol{\sigma}^a\}) &= \sum_{\boldsymbol{\xi}} P(\beta,\boldsymbol{\xi}|\{\boldsymbol{\sigma}^a\}) = \frac{\sum_{\boldsymbol{\xi}} P(\{\boldsymbol{\sigma}^a\}|\boldsymbol{\xi},\beta)P_0(\boldsymbol{\xi},\beta)}{\int d\beta \sum_{\boldsymbol{\xi}} P(\{\boldsymbol{\sigma}^a\}|\boldsymbol{\xi},\beta)P_0(\boldsymbol{\xi},\beta)} \\
&= \frac{1}{Z(\{\boldsymbol{\sigma}^a\})}\sum_{\boldsymbol{\xi}} e^{-NM\ln\left(2\cosh(\beta/\sqrt{N})\right)} \prod_a \cosh\left(\frac{\beta}{\sqrt{N}}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\sigma}^a\right) \\
&\propto e^{-M\frac{\beta^2}{2}} Z(\beta,\{\boldsymbol{\sigma}^a\}),
\end{aligned}
\tag{9}
$$

Iterative equation for temperature prediction:

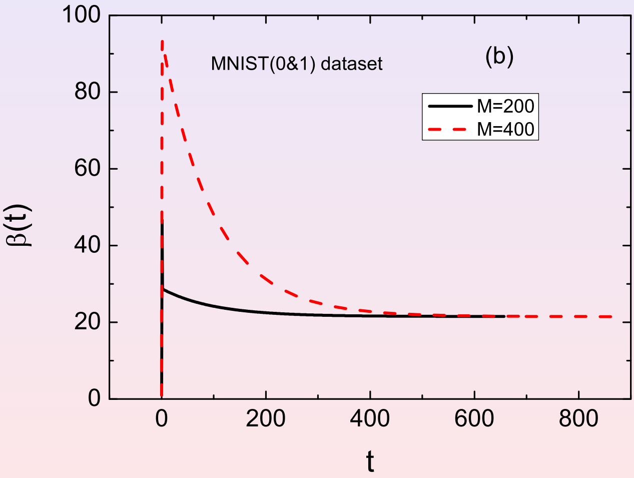
$$
\frac{\partial \ln Z(\beta,\{\boldsymbol{\sigma}^a\})}{\partial\beta} = N\alpha\beta. \tag{10}
$$

Nishimori condition (physics) vs. EM algorithm (statistics)!

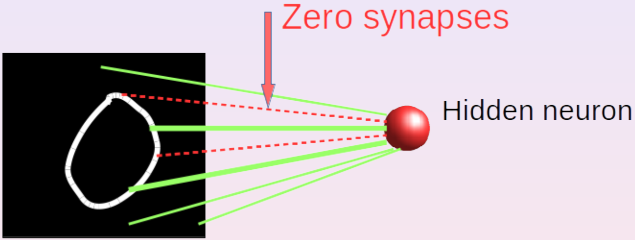
Test on synthetic data



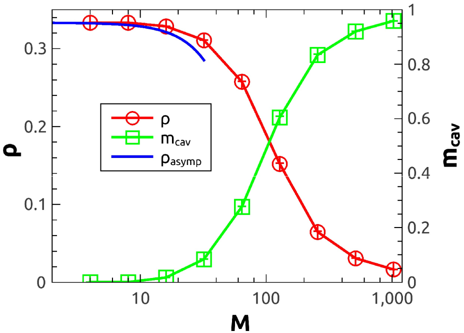
Test on (0,1) handwriten digits

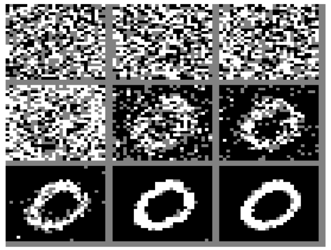


Zero synapses: concept formation



Zero synapses: concept formation





ρ—sparsity level of synapses: $m_{\rm cav}$—overall strength of messages

Huang, arXiv:1703.07943

Summary

- Data determines the weight uncertainty.
- (dis)continuous phase transitions revealed.
- Easy-hard-easy unsupervised learning pattern discovered.
- A quantitative measure of how cold a dataset is provided.
- Role of zero synapses revealed.

Acknowledgements

Supported by the Japan Brain/MINDS project.

Thanks for your attention