Efficient estimation of the Partition Function of Restricted Boltzmann Machines via Annealed Importance Sampling

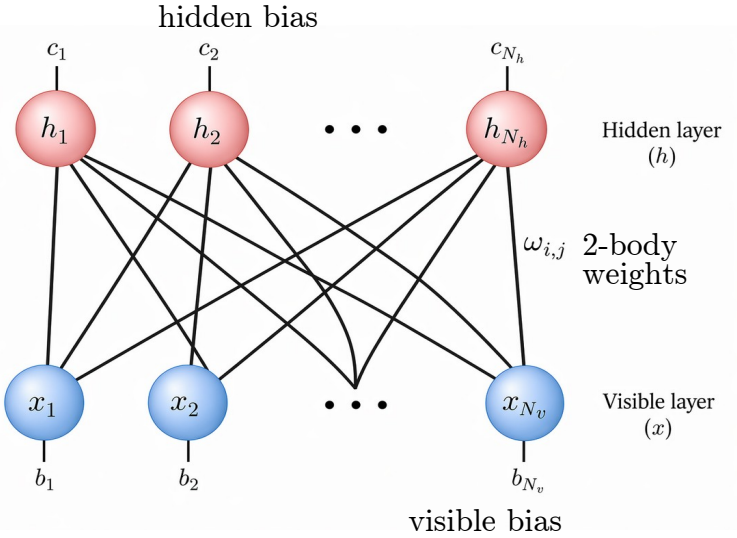
F.Mazzanti[(1)], E.Romero Merino[(2)], A.Prat-Pou[(1)], J.Martí[(1)]
[(1)] Physics Department, Universitat Politècnica de Catalunya, Barcelona
[(2)] Department of Computer Sciences, Universitat Politècnica de Catalunya, Barcelona

The Restricted Boltzmann Machine (RBM)

A constraint version of the Boltzmann Machine neural network

Two sets of nodes: visible **x** (data) and hidden **h** (latent)

Bipartite model: **only connection between different layers are allowed**



A probabilistic and generative Energy Based model

$$E(\mathbf{x}, h) = -\mathbf{x}' \cdot \mathbf{b} - \mathbf{c}' \cdot \mathbf{h} - \mathbf{x}' \cdot \mathbf{W} \cdot \mathbf{h}$$

$$p(\mathbf{x}) = \frac{e^{-F(\mathbf{x})/T}}{Z} \quad , \quad e^{-F(\mathbf{x})/T} = \sum_{\mathbf{h}} e^{-E(\mathbf{x},\mathbf{h})/T} \quad , \quad Z = \sum_{\mathbf{x}} e^{-F(\mathbf{x})/T}$$

Free Energy — **Partition Function**

Its topology implies **conditional independence of variables in the same layer** (nice for sampling)

$$P(\mathbf{h} \mid \mathbf{x}) = \prod_j P(h_j \mid \mathbf{x}), \quad , \quad P(\mathbf{x} \mid \mathbf{h}) = \prod_i P(x_i \mid \mathbf{h})$$

Getting Z (normalized probabilities) is... **essentially impossible because of extremely large dimensionality** 😔

The Restricted Boltzmann Machine (RBM)

The Restricted Boltzmann Machine (RBM)

LETTER Communicated by Yann Le Cun

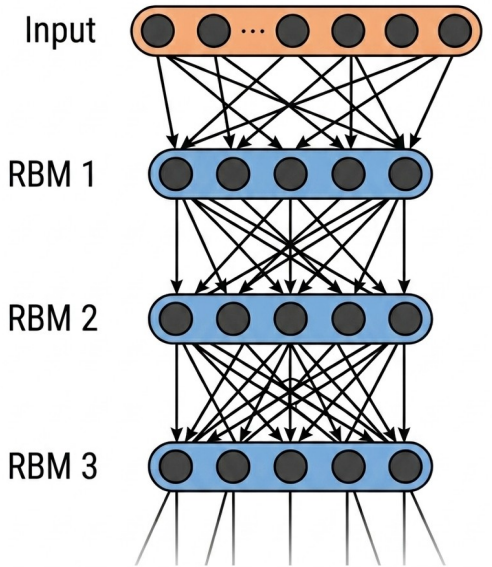
Neural Computation **18, 1527–1554** (2006) © 2006 Massachusetts Institute of Technology

A Fast Learning Algorithm for Deep Belief Nets

Geoffrey E. Hinton
hinton@cs.toronto.edu
Simon Osindero
osindero@cs.toronto.edu
Department of Computer Science, University of Toronto, Toronto, Canada M5S 3G4

Yee-Whye Teh
tehyw@comp.nus.edu.sg
Department of Computer Science, National University of Singapore, Singapore 117543

We show how to use "complementary priors" to eliminate the explaining-away effects that make inference difficult in densely connected belief nets



The Restricted Boltzmann Machine (RBM)

LETTER Communicated by Yann Le Cun

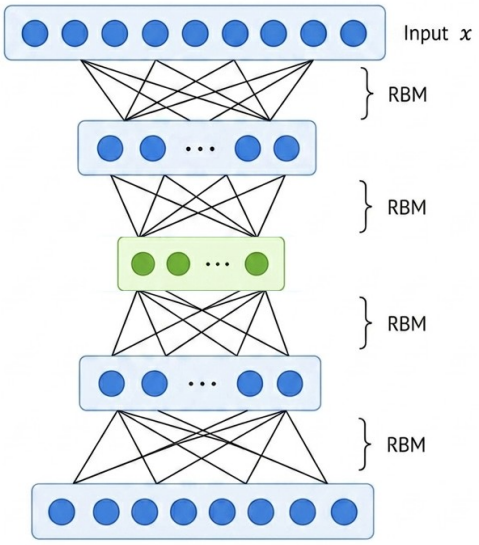
Neural Computation **18, 1527–1554** (2006) © 2006 Massachusetts Institute of Technology

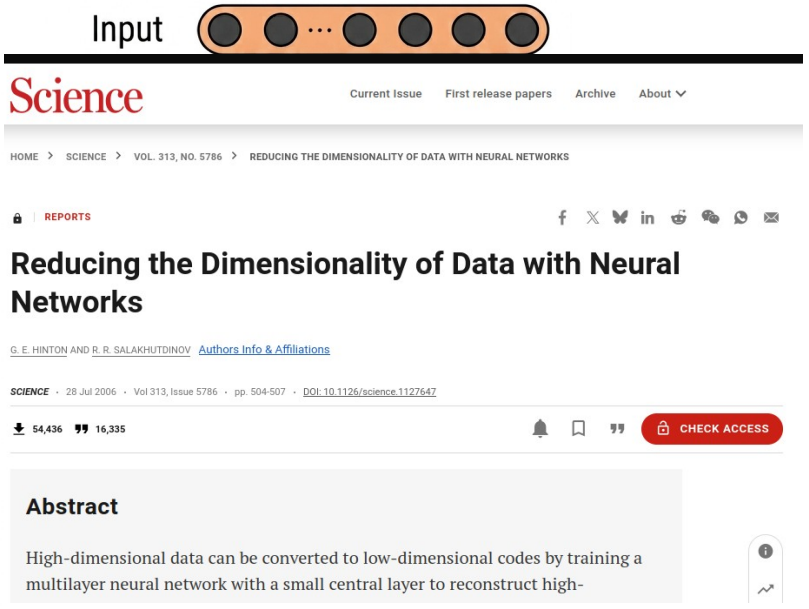
A Fast Learning Algorithm for Deep Belief Nets

Geoffrey E. Hintor
hinton@cs.toronto.edu
Simon Osindero
osindero@cs.toronto.ec
Department of Compu

Yee-Whye Teh
tehyw@comp.nus.edu.
Department of Compu
Singapore 117543

We show how to us
away effects that m





The Restricted Boltzmann Machine (RBM)

LETTER Communicated by Yann Le Cun

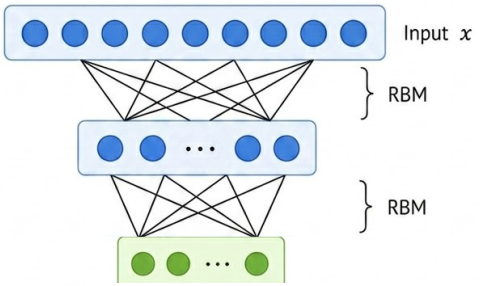
Neural Computation **18, 1527–1554** (2006) © 2006 Massachusetts Institute of Technology

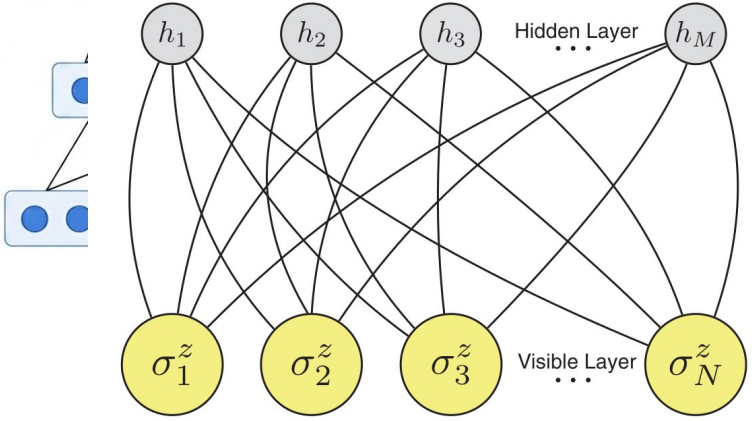
A Fast Learning Algorithm for Deep Belief Nets

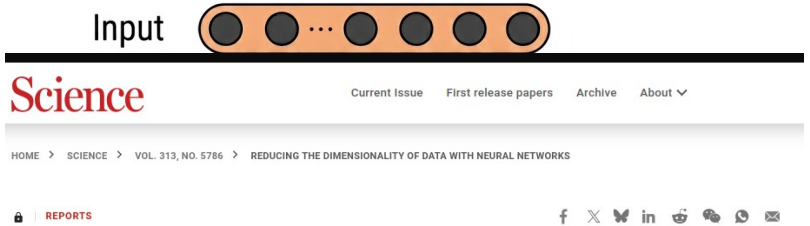
Geoffrey E. Hintor
hinton@cs.toronto.edu
Simon Osindero
osindero@cs.toronto.ec
Department of Compu

Yee-Whye Teh
tehyw@comp.nus.edu.
Department of Compu
Singapore 117543

We show how to us
away effects that m







RESEARCH

Carleo *et al.*, *Science* **355**, 602–606 (2017) 10 February 2017

RESEARCH ARTICLE

MANY-BODY PHYSICS

Solving the quantum many-body problem with artificial neural networks

Giuseppe Carleo[1]* and Matthias Troyer[1,2]

The challenge posed by the many-body problem in quantum physics originates from the difficulty of describing the nontrivial correlations encoded in the exponential complexity

The Restricted Boltzmann Machine (RBM)

Why the Partition Function ?

- **Learning Problems:** weight and biases update rules

$$\omega_{ij}^{(n+1)} \leftarrow \omega_{ij}^{(n)} + \epsilon \cdot \Delta\omega_{ij}^{n)}$$

$$\Delta\omega_{ij}^{n)} = \langle x_i h_h \rangle_+ - \langle x_i h_h \rangle_-$$

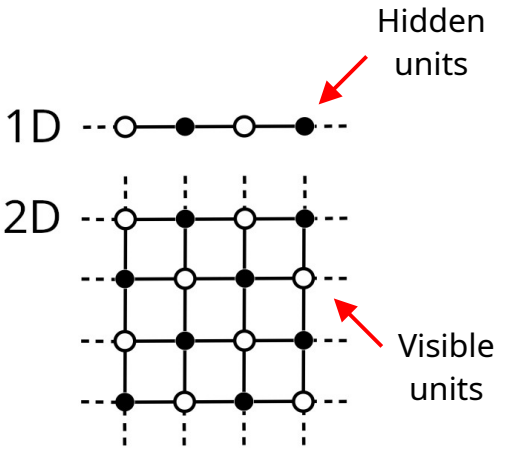
Positive Phase: average over the Training Set

Negative Phase: average over the whole space of $2^{N_v+N_h}$ possible states

$$\langle O \rangle_- = \sum_{\mathbf{x},\mathbf{h}} \frac{e^{-E(\mathbf{x},\mathbf{h})/T}}{Z} O(\mathbf{x},\mathbf{h})$$

- **Statistical Mechanics Applications:** Ex: describe the thermodynamics of magnetic spin systems (Spin Glasses or Ising models) with nearest-neighbor interactions with fixed weights & biases

$$E = -T^2 \frac{\partial}{\partial T} \log(Z) \quad , \quad M_j = \langle x_j \rangle = T \frac{\partial}{\partial b_j} \log(Z)$$



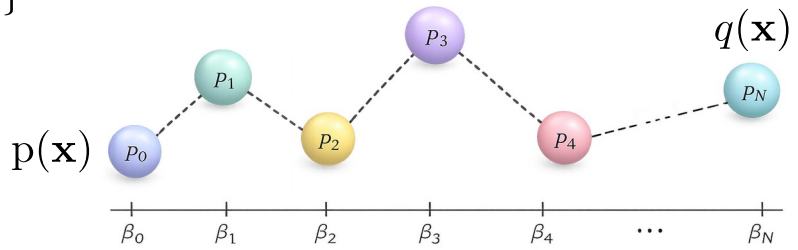
The Annealed Importance Sampling algorithm

R.M. Neal, *Annealed Importance Sampling*, Statistics and Computing 11, 125–139 (2001).

Allows **sampling a very complex probability distribution** $q(\mathbf{x})$ **starting from a trivial to sample one** $p(\mathbf{x})$

Builds a **chain of intermediate probability distributions** $\{p_k(\mathbf{x})\}$ that transforms $p(\mathbf{x})$ into $q(\mathbf{x})$. Define a **sequence of inverse temperatures**

$$0 = \beta_0 < \beta_1 < \beta_2 < \cdots < \beta_N = 1$$



and the corresponding sequence as

$$p_k(\mathbf{x}) = p_0(\mathbf{x})^{1-\beta_k} p_N(\mathbf{x})^{\beta_k} \quad \text{with} \quad p_0(\mathbf{x}) = p(\mathbf{x}) \;,\; p_N(\mathbf{x}) = q(\mathbf{x})$$

which act as an annealing chain that runs from $T = \infty$ to $T = 1$

Each $p_k(\mathbf{x})$ is used as an Importance Sampler of $p_{k+1}(\mathbf{x})$. Also **estimates the ratio of Partition Functions**

$$\frac{Z_q}{Z_p} \approx \frac{1}{M} \sum_{i=1}^{M} \omega^{(i)} \qquad \text{where} \qquad w = \frac{f_1(x_0)}{f_0(x_0)} \cdot \frac{f_2(x_1)}{f_1(x_1)} \cdot \frac{f_3(x_2)}{f_2(x_2)} \cdots \frac{f_N(x_{N-1})}{f_{N-1}(x_{N-1})}$$

$$\langle O \rangle_q \approx \frac{\sum_i \omega^{(i)} O_i}{\sum_i \omega^{(i)}}$$

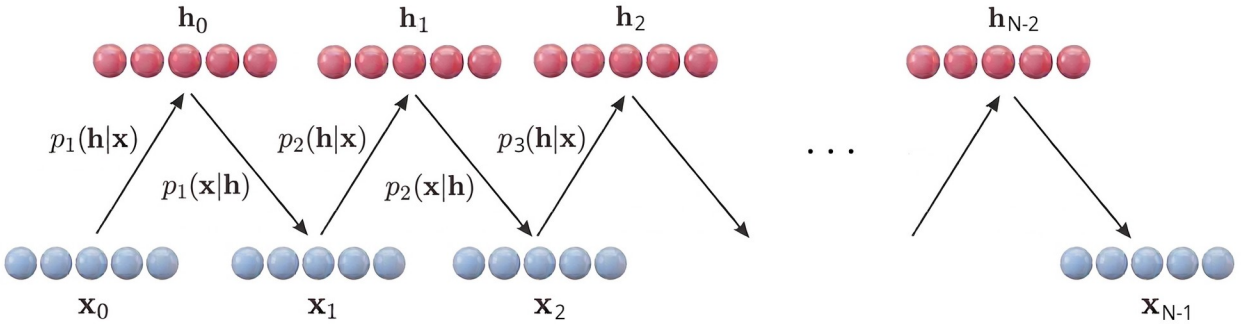
...with $f_n(x) = Z_n p_n(x)$ = unnormalized $p_n(x)$

RBM's bipartite structure allows for efficient Gibbs Sampling of the intermediate AIS distributions.

The Annealed Importance Sampling algorithm

Gibbs sampling starts from a given visible configuration **x** and generates a latent configuration **h** from the conditional probabilities $p(\mathbf{h}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{h})$ in a sequence

Sample $\mathbf{x}$ from $p(\mathbf{x})$:



At this point, **one only needs to define $p(\mathbf{x})$ to start the iteration**

Two important requirements:

- **$p(\mathbf{x})$ must be easy to sample**
- **You must know the Partition function Z_p corresponding to $p(\mathbf{x})$**

Standard solution with no more information: **to start from a mean field model** (*product of experts*)

$$p(\mathbf{x}) = \prod_i p(x_i) = \prod_i e^{-B_i x_i / T} = e^{-\mathbf{B}\cdot\mathbf{x}/T}$$

...but **how do we find a *suitable* mean field $\mathbf{B}$?**

The Mean Fields

- **Simplest choice : set** $\mathbf{B} = 0 \quad \rightarrow \quad p(\mathbf{x}) =$ uniform probability distribution
- **Smarter one: compute it from the training data -if you have it.** That is ok when the training data really acquires a significant amount of probability mass (final learning epochs (?))
- ????

Another possibility is to **use the** $\mathbf{B}$ **that produces the closest probability distribution to that of the RBM.**
This has the advantage of not requiring a training dataset.

One can do that by **minimizing the Kullback-Leibler distance between the mean field** $p(x) = \frac{e^{-\mathbf{B}\cdot\mathbf{x}/T}}{Z_0}$ **and the RBM probability distribution** $q(\mathbf{x})$

$$\mathrm{KL}(p\,||\,q) = \sum_{\mathbf{x}} q(\mathbf{x}) \ln\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) \quad \rightarrow \quad \frac{\partial}{\partial B_\alpha}\mathrm{KL}(p\,||\,q) = 0$$

leading to

$$x \in [0,1]: \; B_\alpha = -T \ln\left(\frac{1}{\langle x_\alpha\rangle_q} - 1\right) \quad , \quad x \in [-1,1]: \; B_\alpha = -T \tanh^{-1}(\langle x_\alpha\rangle_q)$$

In both cases, $\langle x_\alpha\rangle_q$ **is the average value of unit α under the RBM probability distribution** $q(\mathbf{x})$
...but we do not know it ! We have to resort to approximations...

Approaching the Optimal Mean Fields

A. Prat Pou, E. Romero, J. Martí, F. mazzanti, Entropy 27(2), 171 (2025)

We use two different strategies to approximate $\langle x_\alpha \rangle_q$

- **PseudoInverse (Pinv) : find the configuration $\{\mathrm{x}, \mathrm{h}\}$ that minimizes the RBM Energy**

$$\nabla_{\mathbf{x}} E(\mathbf{x}, \mathbf{h}) = 0 \ , \ \nabla_{\mathbf{h}} E(\mathbf{x}, \mathbf{h}) = 0 \quad \rightarrow \quad \mathbf{x}_p = -(\mathrm{W}^+)' \mathbf{c}$$

with W^+ **the pseudoinverse of the two-body weights** matrix. Build $\langle x_\alpha \rangle_q$ rounding $\mathbf{x}_p$ to {0,1} or {-1, 1}

- **Signs from h (Signs_h) :** an heuristic that **minimizes the local RBM Energy for fixed hidden h**

Sample **h** uniformly in {0,1} (or {-1,1}). Then minimize the Energy $E(\mathbf{x}, \mathbf{h})$ for that **h**

$$E = -\mathbf{c}' \cdot \mathbf{h} - \mathbf{x}' (\mathbf{b} - \mathrm{W} \cdot \mathbf{h})$$

constant

set : $\begin{cases} x_i = 1 & \text{if } \ b_i + (\mathrm{W} \cdot \mathbf{h})_i > 0 \\ x_i = 0 & \text{otherwise} \end{cases}$

repeat as many times as you want with different **h**'s to generate different **x**'s. Use them to build $\langle x_\alpha \rangle_q$

This takes much less computational effort than evaluating the PseudoInverse W^+

Putting it to the Test: analyzed systems

We (mostly) study systems where $\log(Z)$ can be evaluated *exactly* :

	N_v, N_h	Weights/Bias	Num. of Weights	Units
GWGM	$20, 180$	$\mathcal{N}(\mu, \sigma)$ $\mu \in \mathcal{N}(-10, 10)$ $\sigma \in \mathcal{N}(20, 10)$	100	$[0, 1]$
MNIST-20h	$784, 20$	learnt 500 epochs	1	$[0, 1]$
1DIsing (1,2,3)	$100, 100$	$\mu = -1, \sigma = 2$ $T = \{0.1,\ 0.01,\ 0.001\}$	100	$[-1, 1]$
1DSpinGlass (1,2,3)	$100, 100$	$\mu = -1, \sigma = 2$ $T = \{0.1,\ 0.01,\ 0.001\}$	100	$[-1, 1]$
2DIsing (1,2,3)	$128, 128$	$\mu = -1, \sigma = 2$ $T = \{0.1,\ 0.01,\ 0.001\}$	100	$[-1, 1]$
2DSpinGlass (1,2)	$32, 32$	$\mu = -1, \sigma = 1$ $T = \{10^{-3}, 10^{-5}\}$	50	$[-1, 1]$
MNIST-500h	$784, 500$	learnt 500 epochs	1	$[0, 1]$

Putting it to the Test: analyzed systems

We (mostly) study systems where $\log(Z)$ can be evaluated *exactly* :

	N_v, N_h	Weights/Bias	Num. of Weights	Units
GWGM	$20, 180$	$\mathcal{N}(\mu, \sigma)$ $\mu \in \mathcal{N}(-10, 10)$ $\sigma \in \mathcal{N}(20, 10)$	100	$[0, 1]$
MNIST-20h	$784, 20$	learnt 500 epochs	1	$[0, 1]$
1DIsing (1,2,3)	$100, 100$	$\mu = -1, \sigma = 2$ $T = \{0.1,\ 0.01,\ 0.001\}$	100	$[-1, 1]$
1DSpinGlass (1,2,3)	$100, 100$	$\mu = -1, \sigma = 2$ $T = \{0.1,\ 0.01,\ 0.001\}$	100	$[-1, 1]$
2DIsing (1,2,3)	$128, 128$	$\mu = -1, \sigma = 2$ $T = \{0.1,\ 0.01,\ 0.001\}$	100	$[-1, 1]$
2DSpinGlass (1,2)	$32, 32$	$\mu = -1, \sigma = 1$ $T = \{10^{-3}, 10^{-5}\}$	50	$[-1, 1]$
MNIST-500h	$784, 500$	learnt 500 epochs	1	$[0, 1]$

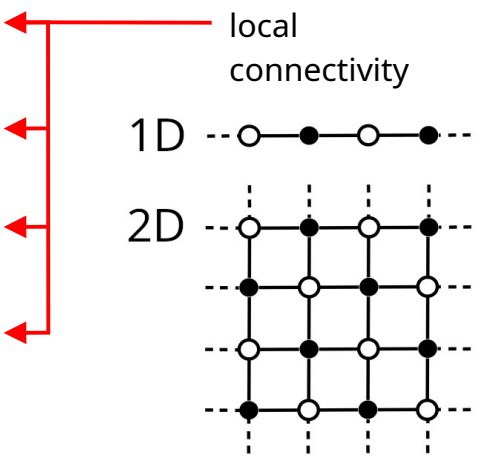
Global connectivity

MNIST dataset of handwritten digits

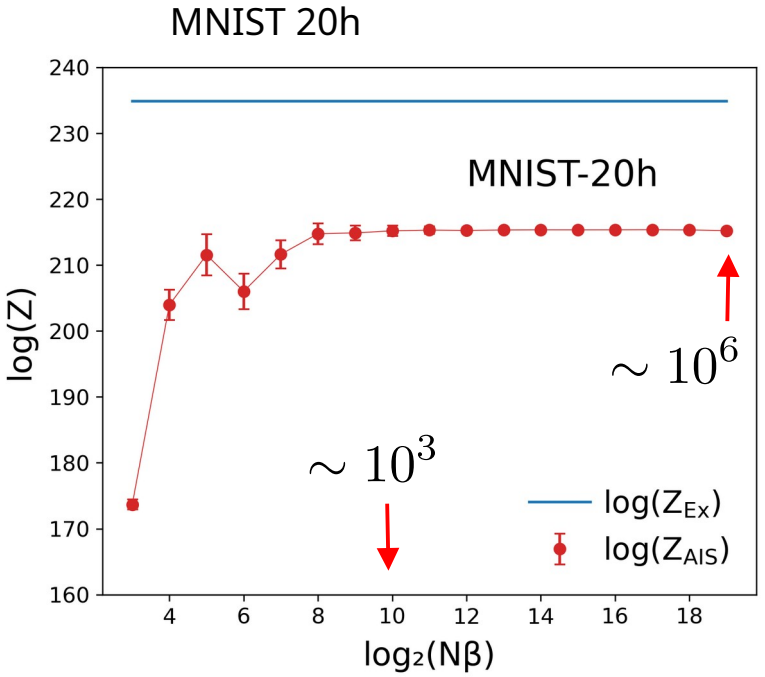
Putting it to the Test: analyzed systems

We (mostly) study systems where $\log(Z)$ can be evaluated *exactly* :

	N_v, N_h	Weights/Bias	Num. of Weights	Units
GWGM	$20, 180$	$\mathcal{N}(\mu, \sigma)$ $\mu \in \mathcal{N}(-10, 10)$ $\sigma \in \mathcal{N}(20, 10)$	100	$[0, 1]$
MNIST-20h	$784, 20$	learnt 500 epochs	1	$[0, 1]$
1DIsing (1,2,3)	$100, 100$	$\mu = -1, \sigma = 2$ $T = \{0.1,\ 0.01,\ 0.001\}$	100	$[-1, 1]$
1DSpinGlass (1,2,3)	$100, 100$	$\mu = -1, \sigma = 2$ $T = \{0.1,\ 0.01,\ 0.001\}$	100	$[-1, 1]$
2DIsing (1,2,3)	$128, 128$	$\mu = -1, \sigma = 2$ $T = \{0.1,\ 0.01,\ 0.001\}$	100	$[-1, 1]$
2DSpinGlass (1,2)	$32, 32$	$\mu = -1, \sigma = 1$ $T = \{10^{-3}, 10^{-5}\}$	50	$[-1, 1]$
MNIST-500h	$784, 500$	learnt 500 epochs	1	$[0, 1]$



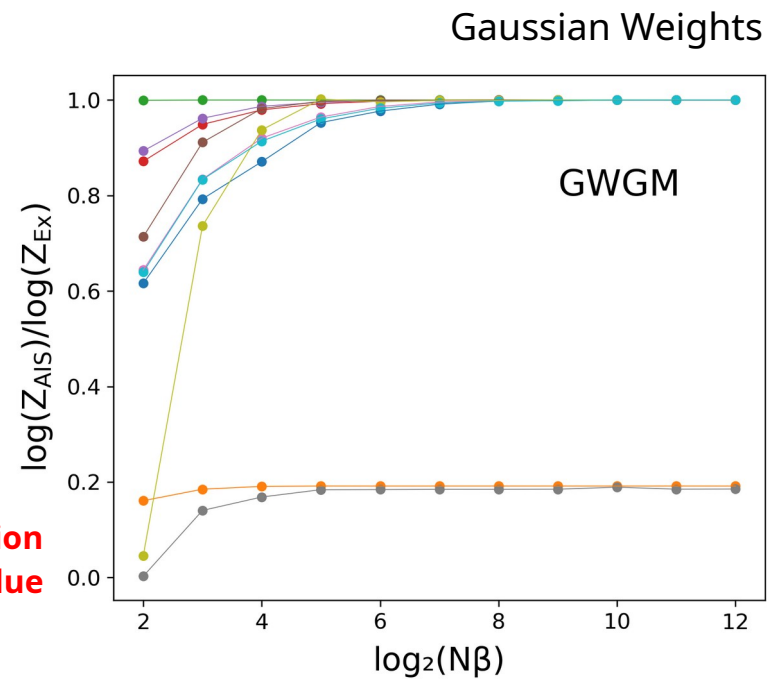
Results of AIS starting from B=0



EXACT RESULT

$N_s = 4096$

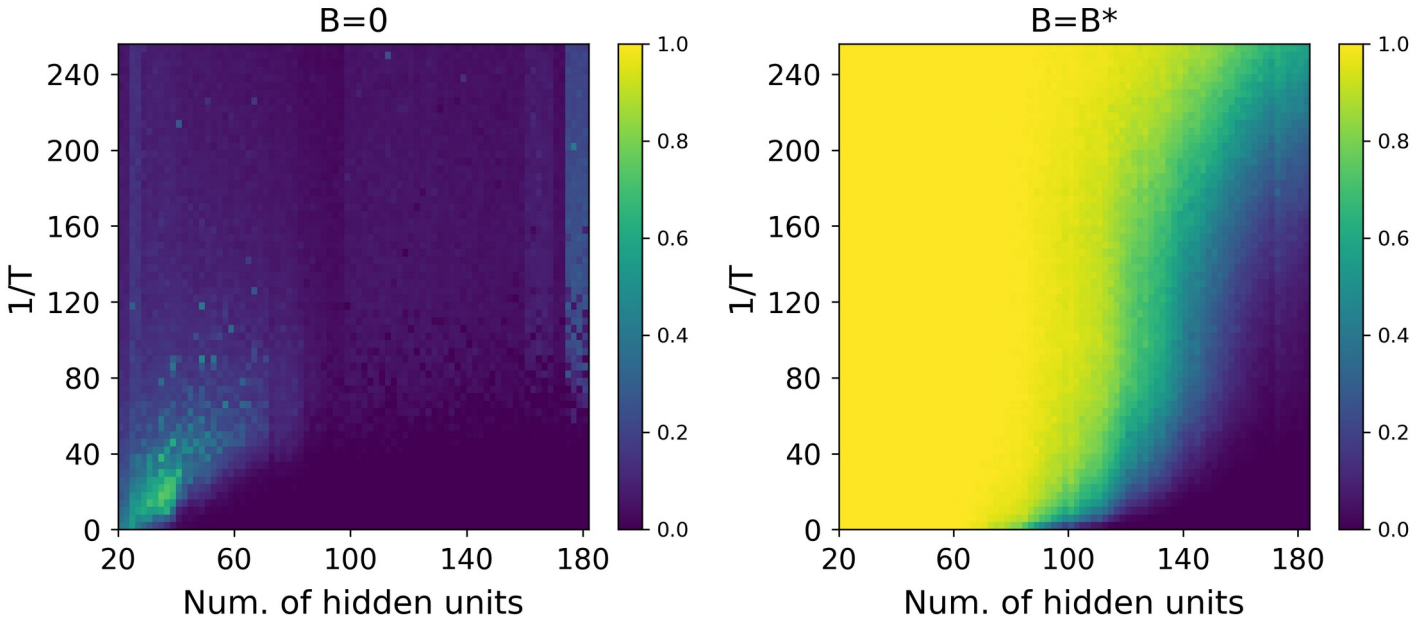
AIS samples



Ratio of the AIS prediction to the exact value

Comparison between B=0 and B=B*

For one selected (hard!) GWGM, as a function of the number of hidden units



B* works better

B* does not always produce good results
(it is always mean field)

B* suffers more with a larger number of hidden units

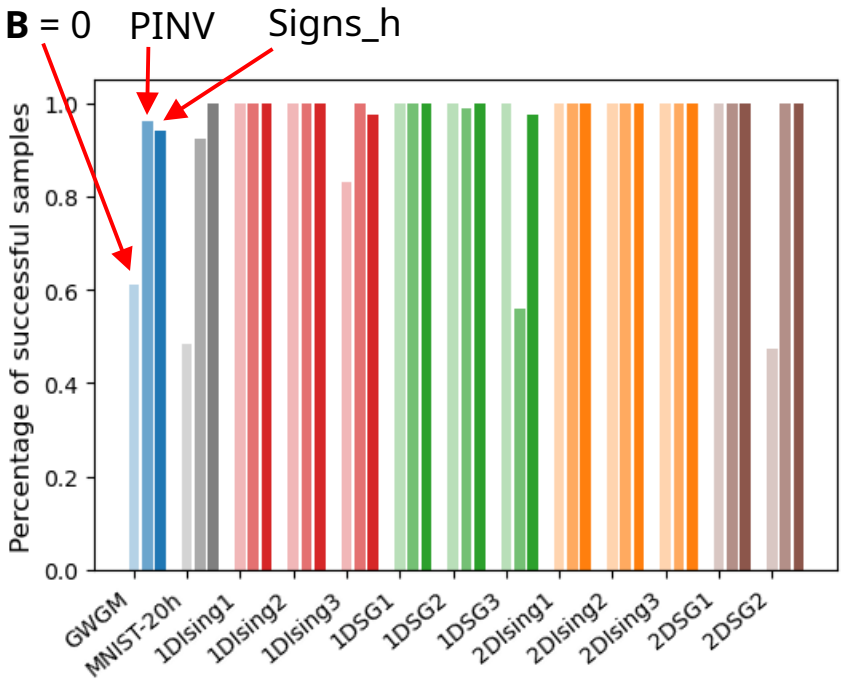
Percentage of AIS samples producing a relative error lower than 5% with respect to the exact $\log(Z)$

Results from the different Mean Fields

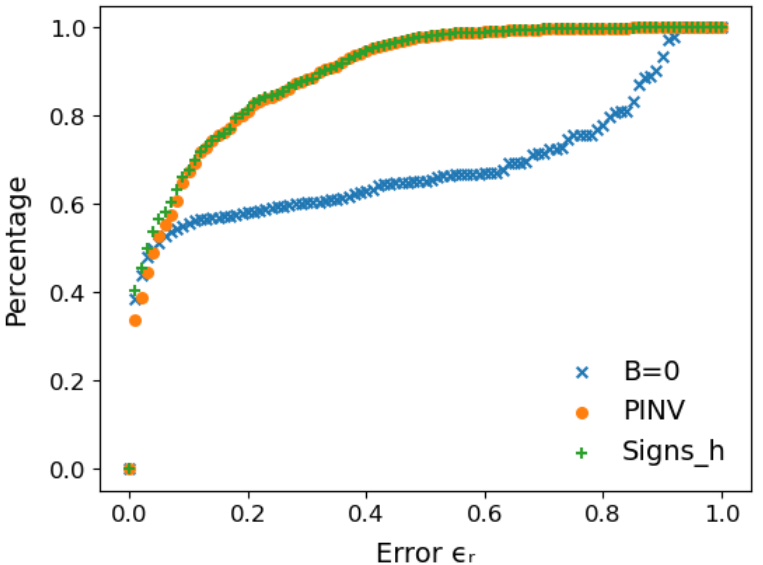
Percentage of AIS samples with a relative error lower than 5% with respect to the exact log(Z) for the different problems analyzed

$N_s = 1024$, $N_\beta = 4096$

AIS samples **Distributions**

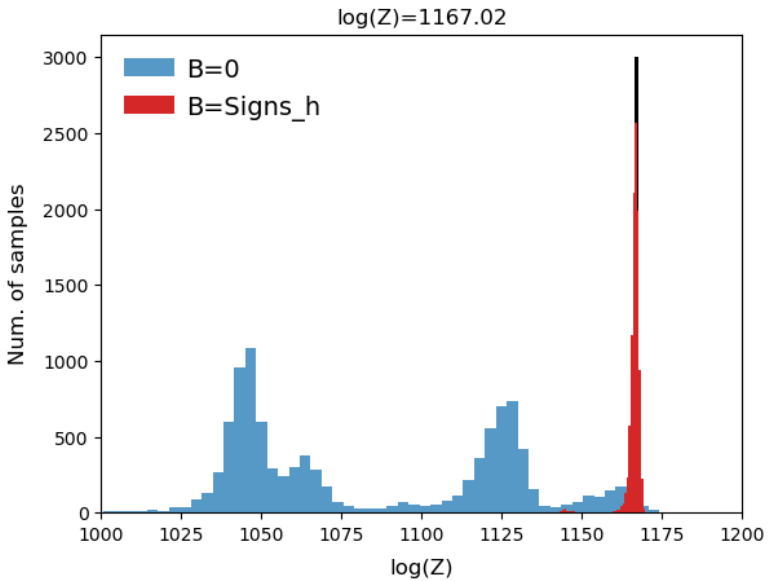


Percentage of AIS samples for GWGM weights with a relative error lower than or equal to ϵ_r with respect to the exact value log(Z)

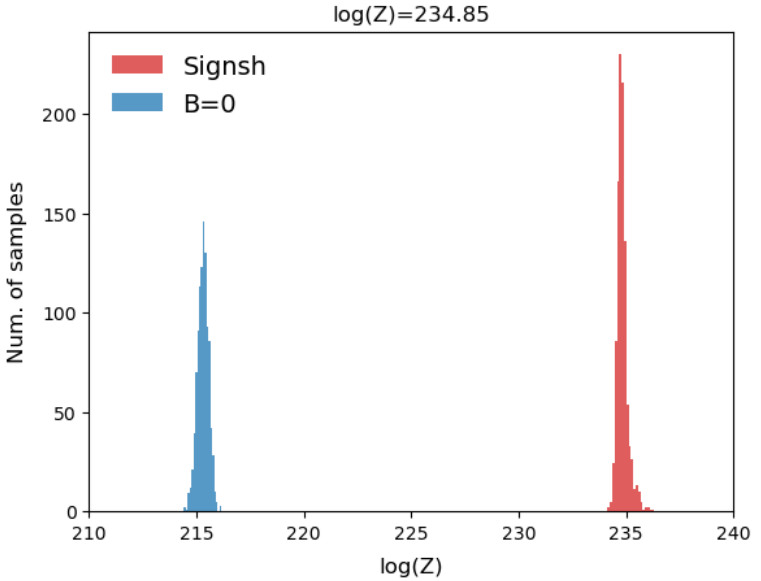


Results from the different Mean Fields

Distribution of samples



MNIST 20h



GWGM

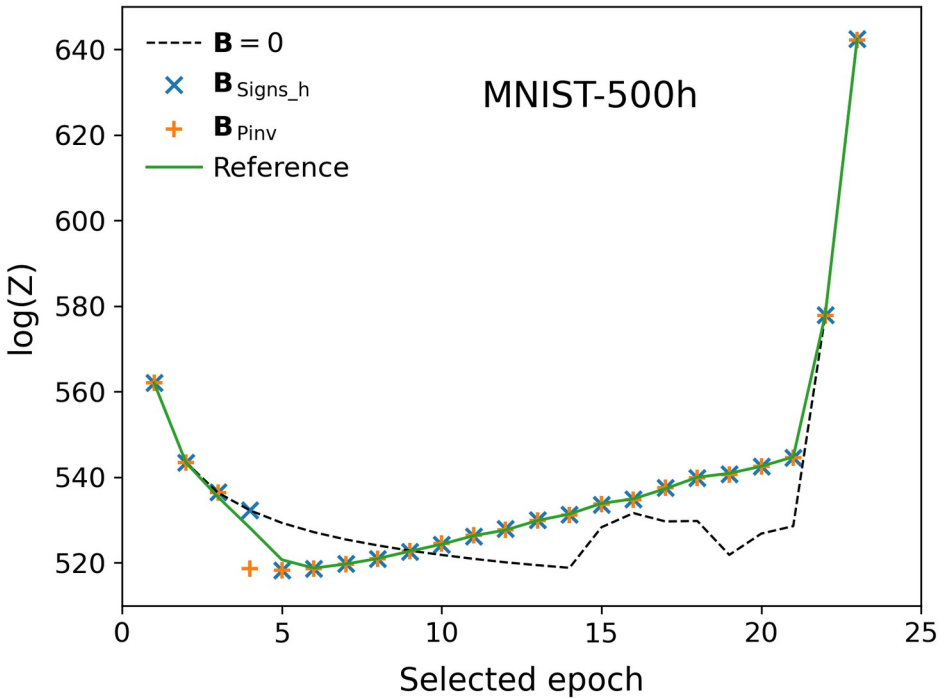
$$Z = \frac{1}{N}\sum_i \omega_i \quad \rightarrow \quad \log(Z) = \log\left(\frac{1}{N}\sum_i e^{\log \omega_i}\right)$$

The result is completely dominated by the largest samples of $\{\log \omega\}$

Comparison to the Standard Reference

Salakhutdinov & Murray,
Proc. 25th Conf. On Machine Learning, 872 (2008)

When a training dataset is available, just use it to estimate $\langle x_\alpha \rangle_q$



$$B_\alpha = -\ln\left(\frac{1}{\langle x_\alpha \rangle_q} - 1\right)$$

Large size system: $N_v = 784$, $N_h = 500$

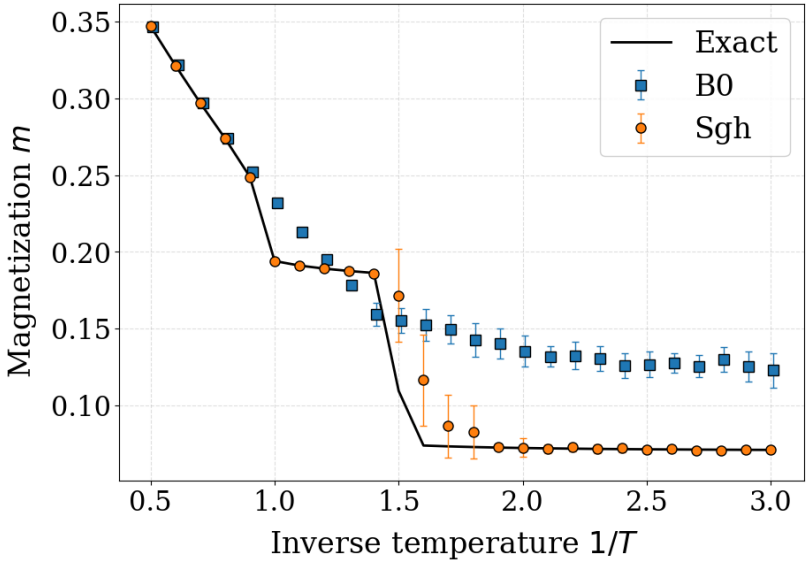
No access to the exact value of $\log(Z)$

Reference computed with $N_\beta = 2^{20}$ intermediate distributions and $N_s = 1024$ AIS samples

The reference line is indistinguishable from the proposed strategies (Sgh, Pinv) and differ from the B=0 prediction -notice we show $\log(Z)$

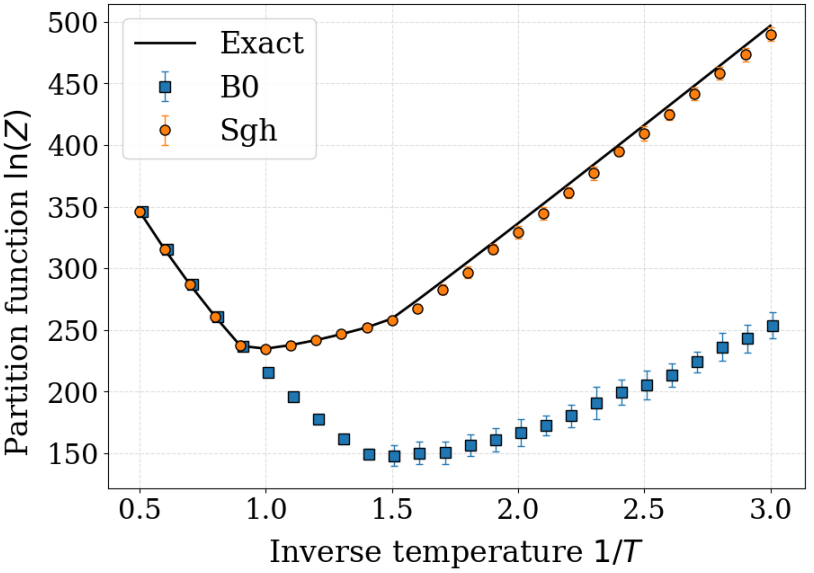
Other Thermodynamic quantities – Phase transitions

work in progress !!



$$\langle O \rangle_q \approx \frac{\sum_i \omega^{(i)} O_i}{\sum_i \omega^{(i)}}$$

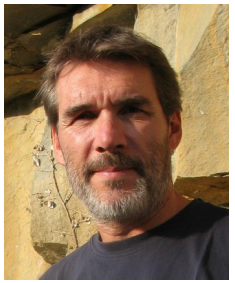
MNIST 20h



Summary and outlook

- AIS is a good tool to estimate Z efficiently if you use the right initialization.
- Mean Field initial probability distributions work well. Even **B = 0** works in many cases, but it is better to use system informed mean fields (depending on the RBM parameters) to approach the Optimal ones with a small budget.
- The proposed initializations perform as the reference strategy employed in learning problems, but they can also be used in the absence of a training set.
- ...and in the end, you don't need to select one or the other. Since AIS samples are obtained from different chains run in parallel, you can simply combine samples starting from all of them and the log-sum-exp operation will favor the best ones.
- Can we use AIS with the proposed strategies to evaluate the negative phase in learning problems ?

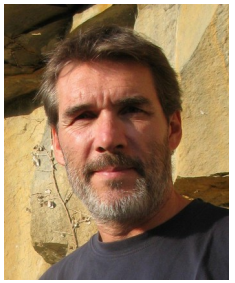
Arnau Prat
Physics Department
Universitat Politècnica de Catalunya



Enrique Romero
Computer Science Department
Universitat Politècnica de Catalunya

Jordi Martí
Physics Department
Universitat Politècnica de Catalunya

Arnau Prat
Physics Department
Universitat Politècnica de Catalunya



Enrique Romero
Computer Science Department
Universitat Politècnica de Catalunya

Jordi Martí
Physics Department
Universitat Politècnica de Catalunya

Thanks for your attention

😊