Stochastic Thermodynamics of Learning

Sebastian Goldt
Institut de Physique Théorique (IPhT), CNRS, CEA, Université Paris-Saclay

At a glance

Background Information processing is constrained by the laws of thermodynamics. For example, erasing a bit requires at least $k_B T \ln 2$ in dissipated heat.

Goal Find the fundamental energetic limits of learning: How much dissipation is necessary to learn?

Results

- The dissipation of any learning device, e.g. a neural network, is an upper bound on the amount of information it can extract from data or learn from a teacher.
- There is a trade-off between dissipation, speed and reliability of any learning device in the steady state.

Perspectives

- Can quantum coherence increase the thermodynamic efficiency of learning?
- Do biological networks, e.g. the retina, show signs of adaptation with respect to thermodynamic constraints?

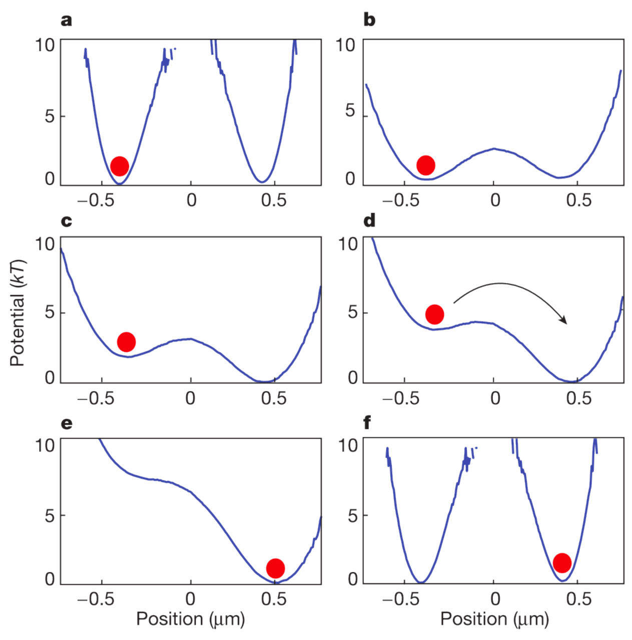
Motivation: the fundamental thermodynamic cost of information processing

Landauer's erasure principle: $W \geq k_B T \ln 2$

Experimentally [1, 2]: overdamped colloidal particle in a laser trap:

$$\dot{x}(t) = -\mu \partial_x V(x, \lambda) + \zeta(t)$$

$$\langle \zeta(t) \zeta(t') \rangle = 2D\delta(t - t') \qquad D = T\mu$$

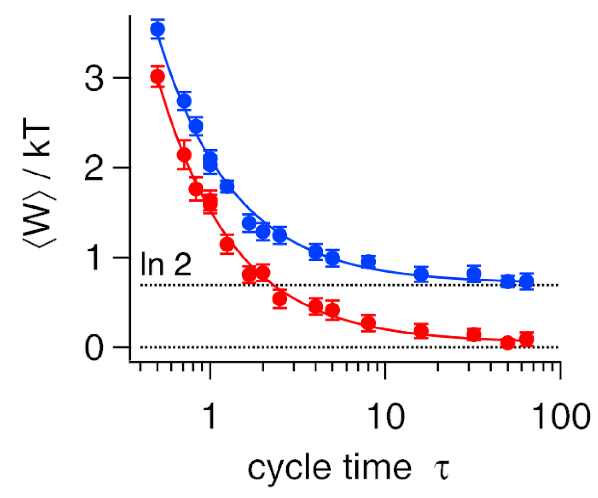


Experimental protocol for the erasure of a single bit, reprinted from [1].

Stochastic Thermodynamics [3] provides consistent definitions of heat and work along single trajectories for small, fluctuating systems far from equilibrium:

$$đw = (\partial_\lambda V(x, \lambda))\,\mathrm{d}\lambda$$

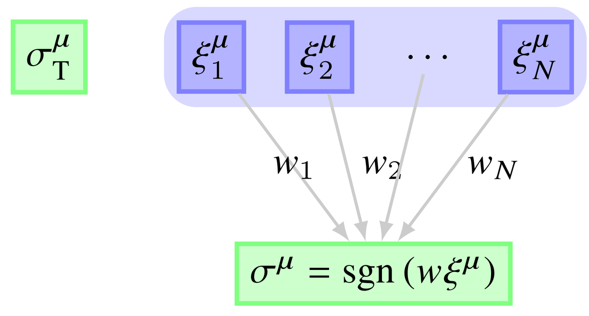
$$đw = \mathrm{d}V + đq$$



Work performed during bit erasure (blue) and using a similar, but symmetric protocol without erasure (red) [2].

The cost of learning?

Toy model Given inputs $\xi^\mu \in \mathbb{R}^N$ with fixed true labels $\sigma_{\mathrm{T}}^\mu = \pm 1$, $\mu = 1, \ldots, P$, a Perceptron with weights $w \in \mathbb{R}^N$ gives outputs $\sigma^\mu = \mathrm{sgn}(w\xi^\mu)$.



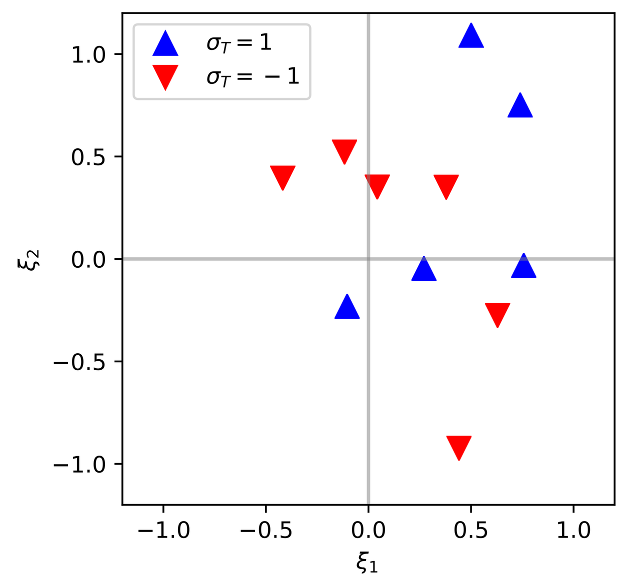
Goal of learning Adjust the weights w s.t. $\sigma_{\mathrm{T}}^\mu = \sigma^\mu$ for as many inputs as possible with minimal dissipation.

Dynamics Langevin equations for w:

$$\dot{w}_t = -w_t + f\left[w_t, \{\xi^\mu, \sigma_{\mathrm{T}}^\mu\}, t\right] + \zeta(t)$$

$$\langle \zeta_n(t) \zeta_m(t') \rangle = 2D\delta_{nm}\delta(t - t')$$

Inferring a model from data



True labels σ_{T}^μ are drawn i.i.d. from $p(\sigma_{\mathrm{T}}^\mu) = 1/2$, independent of ξ^μ and of each other. We can show that for any P, N and learning algorithm with the above dynamics [4]

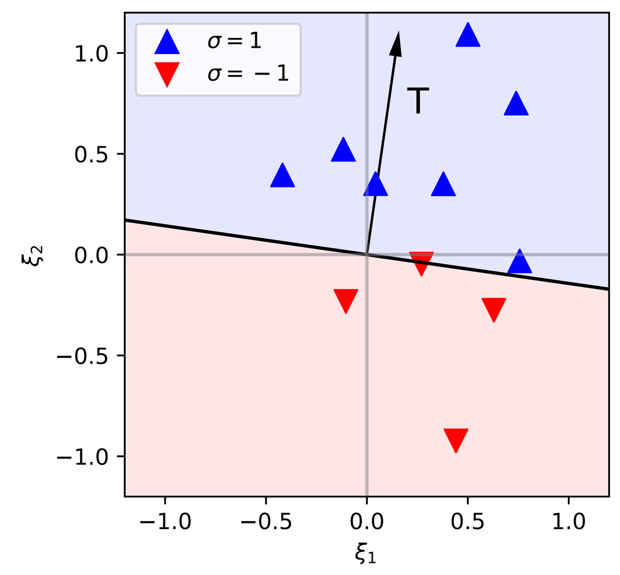
$$\sum_{\mu=1}^{P} I(\sigma_{\mathrm{T}}^\mu : \sigma^\mu) \leq \sum_{n=1}^{N} [\Delta S(w_n) + \Delta Q_n]$$

$I(\sigma_{\mathrm{T}}^\mu : \sigma^\mu)$ mutual information between the true and the predicted label of the μth input

$\Delta S(w_n)$ Change in Shannon entropy of the marginalised distribution $p(w_n)$

ΔQ_n heat dissipated by the nth weight during learning.

The cost of generalising



True labels are now supplied by another Perceptron with weights $T \in \mathbb{R}^N$, the *teacher*, such that $\sigma_{\mathrm{T}}^\mu = \mathrm{sgn}(T\xi^\mu)$. Energetic limits can be formulated by bounding the efficiency of learning [5]

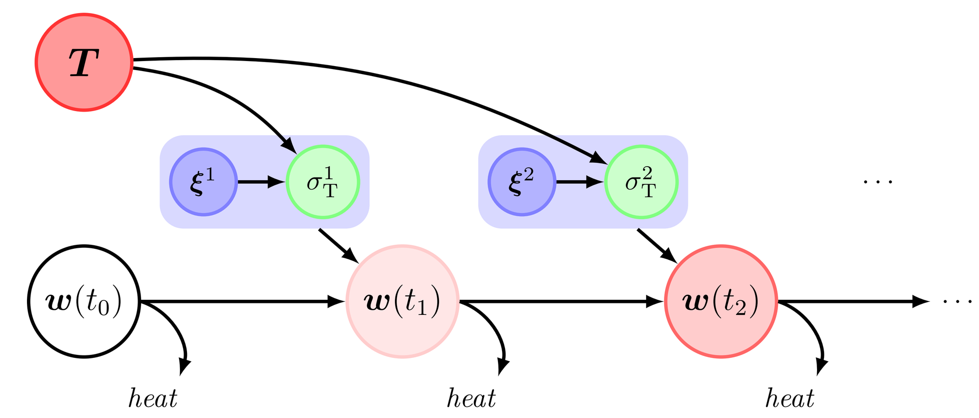
$$\eta \equiv \frac{I(\sigma_{\mathrm{T}} : \sigma)}{\Delta S(w_n) + Q_n} \leq 1$$

$I(\sigma_T : \sigma)$ is the mutual information between the true and the predicted label averaged over ξ; it is related to the generalisation error ϵ_g via

$$I(\sigma_{\mathrm{T}} : \sigma) = \ln 2 - S(\epsilon_g)$$

where $S(x) = -x \ln x - (1 - x) \ln(1 - x)$.

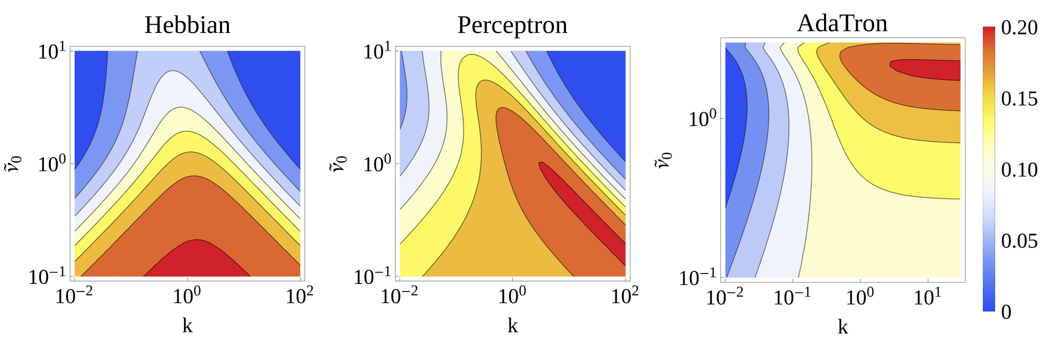
Case study



Learning dynamics with continuous time t:

$$\dot{w}(t) = -kw(t) + \nu(t)\xi^{\mu(t)}\sigma_T^{\mu(t)}\mathcal{F}(\cdot) + \zeta(t)$$

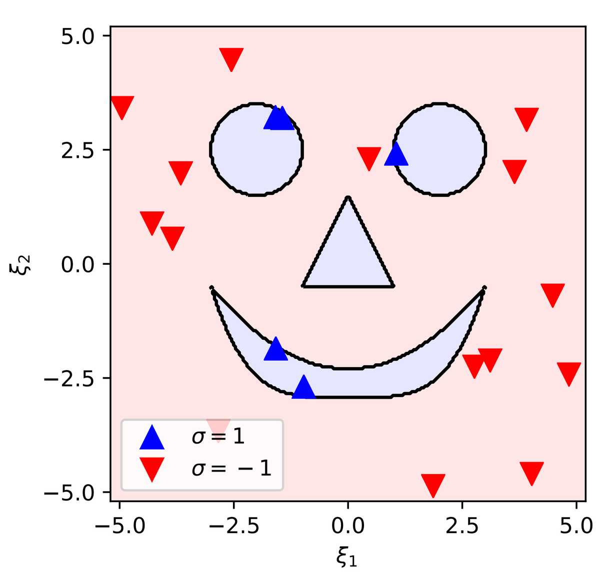
Different learning algorithms can be implemented by choosing the appropriate $\mathcal{F}$, e.g. $\mathcal{F} = 1$ for Hebbian and $\mathcal{F} = \theta(-\sigma_{\mathrm{T}}^\mu w\xi^\mu)$ for Perceptron Learning.



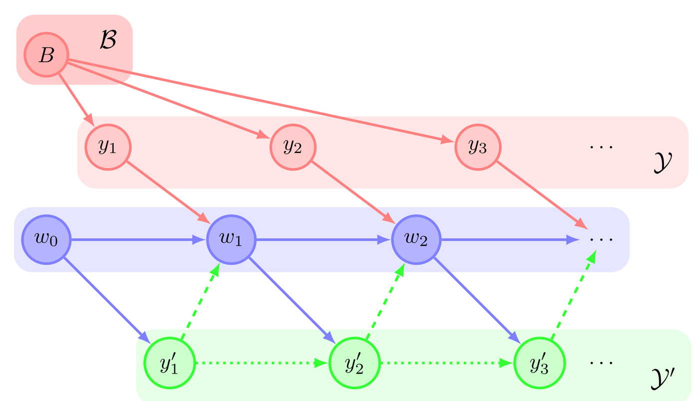
Thermodynamic efficiency of learning η vs learning rate ν and potential stiffness k for online learning by a Perceptron using different learning algorithms [5].

Universal costs of learning and a time-energy-speed trade-off

What about learning more complicated functions, say a smile?
What about deep neural networks? Unsupervised learning? Fluctuations? The role of time?



- Draw samples y from an unknown distribution $q(y|B)$ with possibly time-dependent parameters B.
- The student adjusts the parameters w of his model $p(y|w)$ given the data $\mathcal{Y} = \{y_1, \ldots, y_D\}$, and possibly using feedback from its own outputs $\mathcal{Y}' = \{y'_1, \ldots, y'_D\}$.



We model the learning dynamics using Bayesian networks like above. The integral fluctuation theorem [6] generalises the previous bounds to give the **universal costs of learning**:

$$\langle e^{-\Delta s_w^{\mathrm{tot}} + i(w_\tau : \mathcal{B}, \mathcal{Y}, \mathcal{Y}')} \rangle = 1$$

Small letters denote quantities along a single trajectory.

- Building on the recent "Thermodynamic uncertainty relation" [7, 8]
- *Reliability of learning* $\mathcal{R}$ ≡ inverse variance of acquired information
- Steady state trade-off between $\mathcal{R}$, the speed of learning ν and the energetic cost of the learning device, measured by its entropy production rate $\dot{S}^{\mathrm{tot}}$:

$$\dot{S}^{\mathrm{tot}} \geq \mathcal{R}\nu^2 t$$

Acknowledgements & References

Part of this work was done under the supervision of U. Seifert at the University of Stuttgart. We thank D. Hartich and P. Pietzonka for stimulating discussions. Funding by the European Union's Horizon 2020 Research and Innovation Program 714608-SMiLe is gratefully acknowledged.

[1] A. Bérut et al., Nature **483**, 187 (2012).
[2] Y. Jun, M. Gavrilov, and J. Bechhoefer, Phys. Rev. Lett. **113**, 190601 (2014).
[3] U. Seifert, Rep. Prog. Phys. **75**, 126001 (2012).
[4] SG and U. Seifert, Phys. Rev. Lett. **118**, 010601 (2017).
[5] SG and U. Seifert, New J. Phys. **19**, 113001 (2017).
[6] SG, in preparation (2018).
[7] A. C. Barato and U. Seifert, Phys. Rev. Lett. **114**, 158101 (2015).
[8] T. R. Gingrich, J. M. Horowitz, N. Perunov, and J. L. England, Phys. Rev. Lett. **116**, 120601 (2016).