Classical versus quantum convergence dynamics in energy-based neural networks

Ilona Kulikovskikh[1], **Tarzan Legović**[2]

[1] Samara University, Russia
[2] Institute of Applied Ecology, Oikon Ltd.; Libertas International University; Ruđer Bošković Institute, Croatia

Problem statement

Energy-based neural networks adopt the principles of statistical physics to guarantee their self-regulation. Contrary to probabilistic models, they do not need labels to train and normalization efforts to compute their scores. Shaping the energy function brings substantial power and flexibility to a learning process, which is normally framed as an optimization problem.

The energy landscape for the modern neural network architectures comprises high but thin barriers surrounding shallow local minima. While there is no clear understanding of convergence dynamics produced by classical optimization techniques, quantum annealing which allows tuning the underlying quantum dynamics and has a hardware implementation looks promising.

To deepen the understanding of the potential of quantum tunnelling in energy-based optimization, this study suggests adopting a new paradigm of simulation or "gamification". We analyse the convergence dynamics through building a real (classical) and the alternate (quantum) world, where we can model, observe, and control the behaviour of neurons, communities of neurons, and different types of interactions between them facing open challenges in AI.

Energy landscape

Let us first present a neuron, which maps vectors of n binary inputs $\mathbf{x} \in \{-1, 1\}^n$ to binary outputs $y \in \{-1, 1\}$ through the nonlinear function ℓ, which gives desired outputs $\ell : \boldsymbol{\omega}\mathbf{x} \to \mathrm{l}$, where $\boldsymbol{\omega} \in \mathrm{R}^n$ is the vector of synaptic weights. Both inputs and outputs are independent identically distributed unbiased random variables.

Given p pairs input patterns $\{\mathbf{x}^\mu\}$ and their desired outputs $\{l^\mu\}$, where $\mu = \{1, ..., p\}$, $p = \alpha n$, $\alpha < 0.83$, the problem consists in finding $\boldsymbol{\omega}$ such that all inputs are classified correctly, i.e. $y^\mu l^\mu > 0$. It can be viewed as a search in the weight space for the global minimum of the energy E [2, 4]:

$$E(\boldsymbol{\omega}) = p - \sum_{\mu=1}^{p} \Theta\left(y^\mu l^\mu(\boldsymbol{\omega})\right), \quad 0 \le E \le p, \tag{1}$$

which represents the number of misassociated patterns with the Heaviside step function Θ. The equation (1) defines an n-dimensional landscape which may have many valleys of low energy (local minima) and zero-energy (solutions). The minimization of the energy function $E(\boldsymbol{\omega})$ can be presented as the equation of motion for the overdamped system:

$$\boldsymbol{\omega}(t+1) = \boldsymbol{\omega}(t) - \eta \frac{\partial}{\partial \boldsymbol{\omega}} E(\boldsymbol{\omega}), \tag{2}$$

which is the same as the gradient descent method with the update step t and the step size η.

Quantum effects in energy-based optimization

Let us identify the binary variables ω with physical quantum spins and consider the following time-dependent Hamiltonian [1, 3]:

$$\hat{H}(t) = E(\hat{\boldsymbol{\omega}}) + \frac{1}{2\rho(t)} \hat{\mathbf{p}}^2, \tag{3}$$

where $\hat{\boldsymbol{\omega}}$ denotes degrees of freedom and $\hat{\mathbf{p}}$ is the momentum that satisfies the commutation relation $[\hat{\boldsymbol{\omega}}, \hat{\mathbf{p}}] = i\hbar$, and

$$\hat{\rho}(t) = \frac{1}{Z} \exp(-\beta \hat{H}(t)), \tag{4}$$

where $\hat{\rho}(t)$ represents the mass of the weights and increases over time throughout the quantum annealing process, where the quantum fluctuations appear at $\beta \to \infty$. The average of the logarithm of the partition function

$$Z = \mathrm{Tr}(\exp(-\beta \hat{H}(t)))$$

can be analysed with the Suzuki–Trotter decomposition, which leads to the study of a classical Hamiltonian acting on a system of M interacting Trotter replicas of the original classical system coupled in an extra dimension $\hat{\boldsymbol{\omega}}_k$, where $k = \{1, ..., M\}$. The replicated system needs to be studied in the limit $M \to \infty$ to recover the path integral continuous quantum limit and to make the connection with the behaviour of quantum devices.

The optimization process (2) deals with many replicated realizations or paths $\boldsymbol{\omega}_k(t)$ with the index k (imaginary time). Many realizations are urged into the solution w^* when (4) takes relatively large values. The control over the value of the function $\hat{\rho}(t)$ allows regulating the quantum dynamics to ensure more flexible optimization. When $\rho(t) \to \infty$, (3) recovers the classical optimization problem.

LIGHT neuron

We introduce a new function ℓ, which we refer to as LIGHT (**L**og**I**stic **G**rowth with **H**arves**T**ing), to model inner processes inside neurons. It enriches the state-of-the-art perspective of how neurons receive electrical impulses from other cells, accumulate them, and generate an action potential spike if a threshold value is exceeded. The population of impulses starts growing with the rate r by the logistic law. After time T, it is harvested with the rate E. The sizes of population at 0 and T are specified.

For any time $t \in \mathrm{R}$, harvest time instant $T > 0$, *per capita* growth rate $r > 0$ and *per capita* harvest rate $E \ge 0$, a population of impulses inside a neuron $\ell^{r,E}(t)$, is such that

$$\lim_{t \to -\infty} \ell^{r,E}(t) = 0, \ \lim_{t \to \infty} \ell^{r,E}(t) = \varepsilon,$$

where ε is the extent to which r is impacted by E, develops according to:

$$\ell^{r,E}(t) = \begin{cases} \varepsilon \mathrm{e}^{\left(\ln_{\boldsymbol{q}} N_T + \frac{E}{r}\right)\mathrm{e}^{-r(t-T)}} & \text{if } t \ge T, \\ \varepsilon \mathrm{e}^{(\ln_{\boldsymbol{q}} N_0)\mathrm{e}^{-rt}} & \text{otherwise,} \end{cases} \tag{5}$$

where an initial population size is $N_0 = \ell^{r,E}(0)$, a population size at T is $N_T = \ell^{r,E}(T)$, and $\ln_{\boldsymbol{q}}(x)$ is the *q*-logarithm, where *q* is the rate with which population grows when smaller.
The parameter q generalizes the Verhulst ($q = 1$) and Gompertz ($q \to 0$) laws of population dynamics. If $q = 1$, $T = 0$, $N0 = 0.5$, $r = 1$, and $E = 0$, the function reduces to the sigmoid.

Quantum dynamics in LIGHTs

Interacting replicas represents the quantum effect. By changing the magnitude of $\rho(t)$ for each replica $\boldsymbol{\omega}_k(t)$, we can shift from a sharp local minimum to a wider local minimum preserving better convergence dynamics in the energy-based models. Figure 1 depicts different modes of regulating the traces of weight replicas.

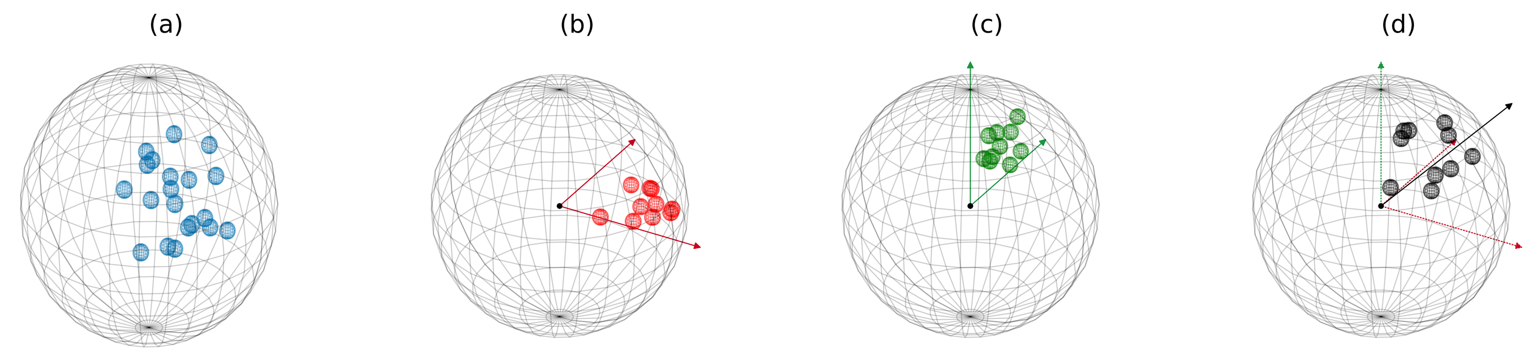

Figure 1. Regulating weight replicas through impulses inside the LIGHT neuron: (a) no regulation; (b) regulation by growth ; (c) regulation by harvesting; (d) regulation by growth and harvesting

Experimental design

A set of synthetic datasets ($m = 1000$, $n = 2$) were generated and randomly split into training ($\alpha = 0.8$) and testing ($1 - \alpha$) subsets. To examine quantum effects, we implemented one sigmoid neuron which does not regulate the dynamics (see Figure 1 (a)) and the LIGHT neuron which does (see Figure 1 (d)). To optimize the weights, we used Stochastic Gradient Descent (SGD), in line with its adaptive modifications (Adam and Adagrad), with the default parameters and batch size $|B(t)| = 75$, which were distributed to each Trotter slice k. Figure 2 summarises accuracy values computed with regard to the energy function for each subset. For better performance, the curves must lie above the diagonal.

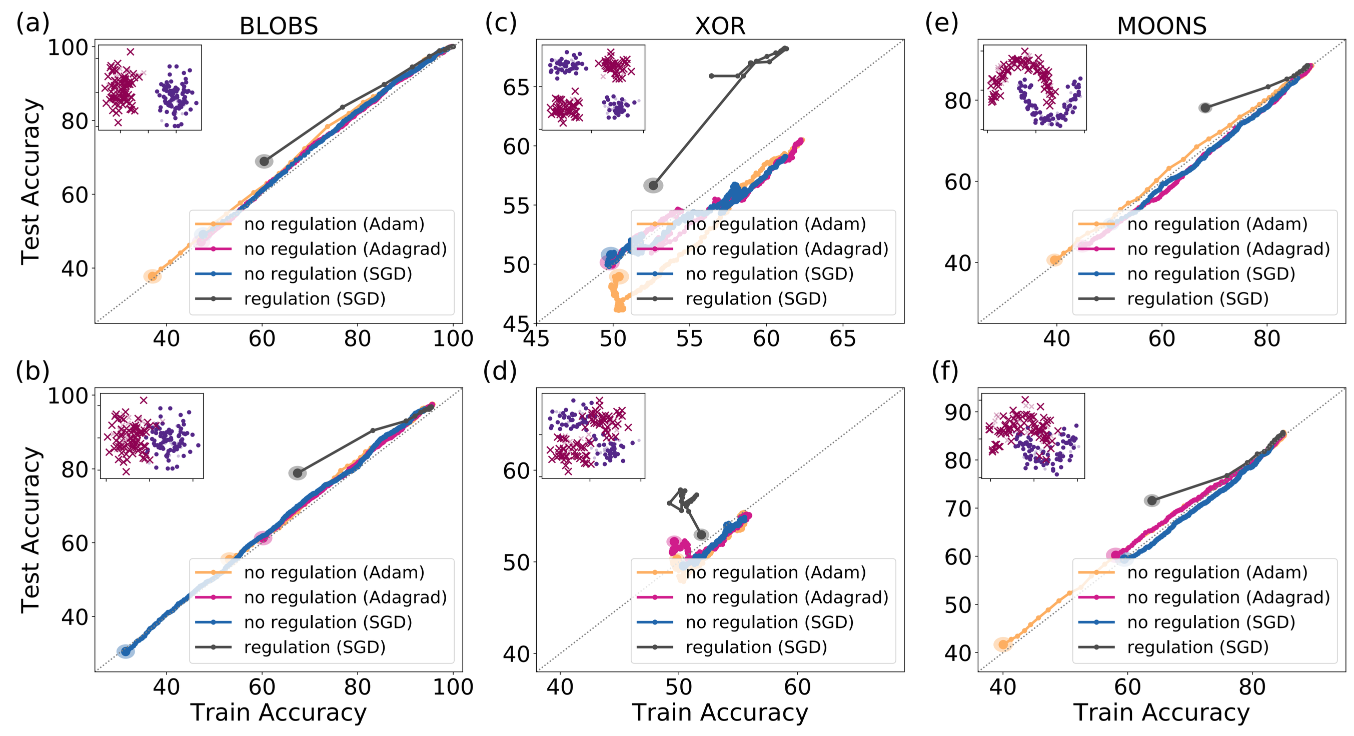

Figure 2. Test/train accuracy curves over epochs on synthetic datasets, averaged over $n_{\mathrm{run}} = 10$

As we can see, SGD equipped with controllable quantum effects outperforms the other methods.

Classical world vs Quantum world

To analyse the convergence dynamics in more details for different network architectures and modes, we use a new paradigm of simulation and build a quantum world which reduces to the classical one under the certain conditions:

Inhabitants. Intelligent inhabitants of the world are the LIGHT neurons. Each LIGHT represents a single-species population of impulses. The environment provides the population with limited resources.

Habitat. A single-species population of the LIGHT neurons rarely lives in isolation from populations of other species, sharing a habitat. The habitat comprises several lands which provide necessary resources for population growth and harvesting. The number of species occupying the same habitat form a community. Their relative abundance is referred to as LIGHT's diversity.

Community. LIGHTs interact with each other and compete for the same resources. The interactions between these populations play a major role in regulating population growth and abundance. Different ways of how the population competes for resources, which balances its growth and harvesting rates, define different species. The maximum size of the population which can be supported by the environment is called carrying capacity. LIGHTs move from one habitat to another. The habitat, which provides enough resources for a population to grow within a long time is called the *global optimum*. A *local optimum* is a temporary habitat, where the population can successfully grow for a short period of time.

Main values of the community:

1. How fast a population comes to the condition when it has enough resources to exist with high competition.
2. How many LIGHTs survive while balancing between lower and higher competition for the resources.

These values allow observing convergence dynamics with quantum effects by controlling the behaviour of neurons, communities of neurons, and different types of interactions between them.

References

[1] C. Baldassi and R. Zecchina. "Efficiency of quantum vs. classical annealing in nonconvex learning problems". In: *Proceedings of the National Academy of Sciences* 115.7 (2018), pp. 1457–1462. DOI: http://dx.doi.org/10.1073/pnas.1711456115.

[2] J. F. Fontanari and R. Köberle. "Landscape statistics of the binary perceptron". In: *J. Phys. France* 51 (1990), pp. 1403–1413. DOI: http://dx.doi.org/10.1051/jphys:0199000510130140300.

[3] M. Ohzeki et al. "Optimization of neural networks via finite-value quantum fluctuations". In: *Scientific Reports* 8.9950 (2018). DOI: http://dx.doi.org/10.1038/s41598-018-28212-4.

[4] T. L. H. Watkin et al. "Learning multi-class classification problems". In: *J. Phys. I France* 2 (1992), pp. 167–180.