STOP LOOKING FOR BIO-INSPIRATION. LET MACHINES HAVE THEIR OWN WORLD

Ilona Kulikovskikh
Information Systems and
Technologies Department
Samara University
Samara 443086, Russia
kulikoskikh.im@ssau.ru

Tarzan Legović
Institute of Applied Ecology, Oikon Ltd.
Libertas International University
Ruđer Bošković Institute
Zagreb 10000, Croatia
tlegovic@oikon.hr

ABSTRACT

Machines must be clearly and fully explainable to be socially acceptable. This is because they operate in areas of high societal significance such as health, police, mobility, or education. However, they still lack the transparency that is vital for their adoption. In an attempt to explain machines' behavior, it is natural to look for bio-inspiration originating in humans and living organisms. Even though machine intelligence still can barely compete with human intelligence, this analogy leads to an unreasonable fear of technological singularity. Before being established, our world has undergone long-term evolution. Observing it through many generations, we still can not explain all mysteries it holds. Since we need machines to complement our society, would it be better to build a new alternative world with clearly defined principles and laws, addressing the current ethical concerns of artificial intelligence? This exhibit suggests developing machine learning projects through building an artificial world. In line with a new interpretation, it brings with it a transparent and enjoyable way of presenting the current advances in the area.

1 THE LANDS OF LIGHT

Let us build an alternate world. We start with a brief introduction of the world setting: inhabitants, habitat, community, and its values. Then, we formulate the convergence rate and generalization error - two fundamental concepts in machine learning - as the key values of the artificial world. Finally, we provide the theoretical justification to show the validity of the principles and laws of the artificial world.

1.1 INHABITANTS

Intelligent inhabitants of the artificial world are the LIGHT (**L**og**I**stic **G**rowth with **H**arves**T**ing) neurons (Kulikovskikh et al., 2019; Kulikovskikh & Legović, 2020; 2021). Each LIGHT represents a single-species population of impulses. The impulses start growing from the initial size at a certain rate. At some arbitrary time instant, they are harvested. The environment provides the population with limited resources.

1.2 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.

1.3 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.

1.4 Main values of the community

The most important value of the community is a *convergence rate*. It defines how fast a population comes to the condition when it has enough resources to exist with high competition.

An increase in a convergence rate changes the behavior of the population with time as follows. Starting with the minimum size, the population grows fast with a large time lag between the beginning of growth and harvesting. When the population is large enough to be supported by a local optimum, it moves to the next habitat. In the end, it comes to the global optimum where the population, close to the carrying capacity, can grow slowly under the same conditions, with the maximum internal competition.

Decreasing the convergence rate results in the opposite behavior. The population starts growing with the minimum size slowly with a short time lag between the beginning of growth and harvesting. When the harvesting occurs, the population, preserving the minimum population size, increases its growth rate. Lower competition for resources at each local minima significantly delays the population's ability to come to the global optimum because it needs more time to exhaust resources at each local optimum.

The transition from a lower convergence rate to a higher convergence rate and vice versa exposes another dominant value of the community - *generalization capability*. It shows the population's ability to survive while balancing between lower and higher competition for the resources.

1.5 Where real and artificial worlds meet

The majority of the LIGHT population speaks the Python language, which they use to communicate with the real world. The sea `class LIGHT()` along with the lands `LIGHT.train()`, `LIGHT.predict()`, `LIGHT.grow()`, and `LIGHT.harvest()` allows LIGHTs effectively augment our reality.

The process of convergence rate optimization on `LIGHT.train()` Land is depicted in Figure 1. Moving from one habitat to another, the community is governed by the two main laws:

Law 1: A higher convergence rate ensures the maximum population with higher competition;

Law 2: A lower convergence rate guarantees the minimum population with lower competition.

The theoretical foundation behind these laws is deferred to Appendix A.

Recent studies reported complex convergence dynamics in optimization methods, which has not been well understood so far. First, overparametrization in deep learning leads to faster convergence (Arora et al., 2018; Li & Liang, 2018; Allen-Zhu et al., 2019; Oymak & Soltanolkotabi, 2019; Liu & Belkin, 2020; Oymak & Soltanolkotabi, 2020; Chen et al., 2021). Second, a step size on testing is usually larger than a step size on training (Bortoli et al., 2020; Li & Arora, 2020; Cohen et al., 2021).

Explaining these findings from the viewpoint of our artificial world, we can conclude that the revealed effect of overparameterization immediately follows from Law 1. Higher generalization capability lowers a convergence rate. By Law 2, this minimizes the population size, which needs a higher growth rate to survive. Following the theoretical implications given in Section A.2, the growth rate is proportional to a step size of an optimizer. This deepens the understanding of the second reported result.

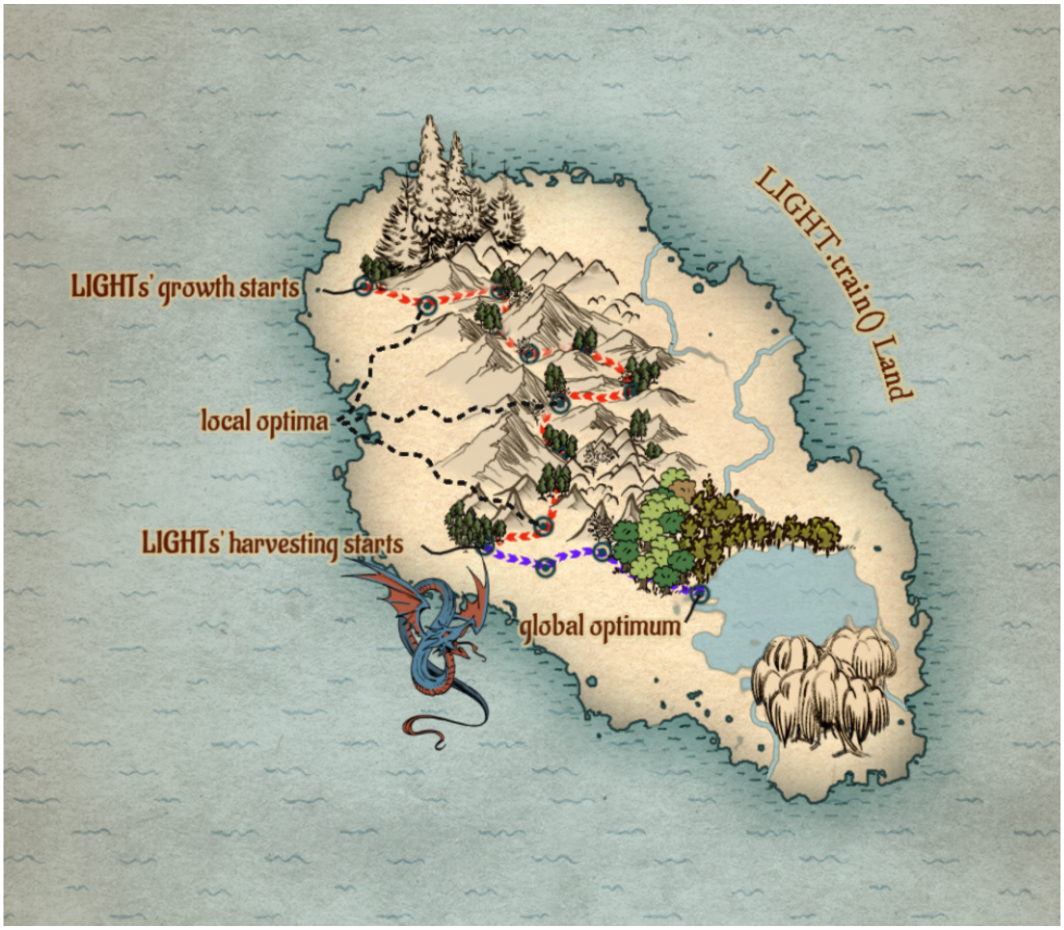

Figure 1: The interpretation of convergence rate optimization on `LIGHT.train()` Land

2 Discussion

While worldbuilding can be presented in the current publishing format (LaTeX + PDF workflows), it complements the submission with new visualization tools, which improve clarity, content delivery, and interactivity. The key point of worldbuilding is to encourage metaphorical thinking, which can spark a better understanding of complex ideas by associating an unfamiliar idea with one that is commonplace. This is one of the fastest ways to build practical intuition around machine learning, addressing the transparency and explainability issues.

Wider adoption of the proposed format can also lead to creating games, such as role-playing games (RPGs), where players take the roles of characters in a fictional setting. RPGs have inherent flexibility and a narrative framework that offers an easy mode for the adoption of machine learning technologies. It promotes collaboration with researchers in other disciplines and creates an accessible learning environment for the younger generation, inviting them to understand, apply and contribute to further development of innovative technologies.

3 Accessibility statement

Worldbuilding accessibility refers to the accessibility of gaming products. Video games can be used not only for entertainment but also for education, rehabilitation, or health, regardless of abilities or age. An increasing number of people interested in them made game accessibility an emerging field of research.

People with disabilities could benefit from the opportunities which video games offer the most. They can acclimatize to technology or use games as an interface between the artificial and real world, where a player, as an avatar, can interact within modern communication systems.

References

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *ICML*, 2019.

S. Arora, N. Cohen, and E. Nazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *ICML*, 2018.

V. De Bortoli, A. Durmus, X. Fontaine, and U. Simsekli. Quantitative propagation of chaos for sgd in wide neural networks. In *NeurIPS*, 2020.

Z. Chen, Y. Cao, D. Zou, and Q. Gu. How much over-parameterization is sufficient to learn deep relu networks? In *ICLR*, 2021.

J. Cohen, S. Kaur, Y. Li, Z. Kolter, and A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *ICLR*, 2021.

W. G. Gray and G. A. Gray (eds.). *Introduction to Environmental Modeling*. Cambridge University Press, Cambridge, UK, 2017.

I. Kulikovskikh and T. Legović. Why to "grow" and "harvest" deep learning models? *CoRR*, abs/2008.03501, 2020. URL `https://arxiv.org/abs/2008.03501`.

I. Kulikovskikh and T. Legović. Painless step size adaptation for SGD. *CoRR*, abs/2102.00853, 2021. URL `https://arxiv.org/abs/2102.00853`.

I. Kulikovskikh, S. Prokhorov, T. Lipić, T. Legović, and T. Šmuc. Biogd: Bio-inspired robust gradient descent. *PLOS ONE*, 14(7):1–19, 07 2019. doi: 10.1371/journal.pone.0219004.

Y. Li and Y. Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *NeurIPS*, 2018.

Z. Li and S. Arora. An exponential learning rate schedule for deep learning. In *ICLR*, 2020.

C. Liu and M. Belkin. Accelerating sgd with momentum for over-parameterized learning. In *ICLR*, 2020.

M. S. Nacson, J. Lee, S. Gunasekar, N. Srebro, and D. Soudry. Convergence of gradient descent on separable data. In *AISTATS*, 2019.

S. Oymak and M. Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *ICML*, 2019.

S. Oymak and M. Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020. doi: 10.1109/JSAIT.2020.2991332.

D. Soudry, E. Hoffer, M.S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *JMLR*, 19:1–57, 2018.

A Theoretical justification for the artificial world

A.1 Problem statement

For a dataset $\{\mathrm{x}_i, y_i\}_{i=1}^m$ with $\mathrm{x}_i \in \mathrm{R}^n$, $y_i \in \{-1, 1\}$, let us minimize an empirical loss function with a weight vector $\theta \in \mathrm{R}^n$:

$$\mathcal{L}(\theta) = \sum_i \ell(y_i \theta^\top \mathrm{x}_i), \tag{1}$$

where ℓ measures the discrepancy between the output y and the model prediction. The gradient descent (GD) finds the weight vector with a fixed step size η:

$$\theta(t+1) = \theta(t) - \eta \nabla_\theta \mathcal{L}(\theta). \tag{2}$$

For a large family of monotone losses with polynomial and exponential tails (Nacson et al., 2019), the derivative of $\ell(t)$ can be presented as $\ell'(t) = -e^{-f(t)}$, where $f(t)$ satisfies $\forall k \in \mathrm{N}$: $\left|\frac{f^{k+1}(t)}{f'(t)}\right| = \mathcal{O}(t^{-k})$. The continuous form of equation 2 ($\eta \to 0$) is equal to $\theta'(t) = \sum_i e^{-f(y_i \mathrm{x}_i^\top \theta(t))} y_i \mathrm{x}_i$, where the weight vector can be presented asymptotically as $\theta(t) = g(t)\hat{\theta} + h(t)$, $h(t) = o(g(t))$, where $g(t)$ defines a convergence rate, $\hat{\theta} = \arg\min_{\theta \in \mathrm{R}^n} \|\theta\|^2$, so that $y_i \theta^\top \mathrm{x}_i \geq 1$ (Soudry et al., 2018; Nacson et al., 2019) . Using $\ell'(t)$, we can write:

$$g'(t)\hat{\theta} = \sum_i e^{-f(g(t) y_i \mathrm{x}_i^\top \hat{\theta} + h(t) y_i \mathrm{x}_i^\top} y_i \mathrm{x}_i \approx e^{-f(g(t)} \sum_i e^{-f'(g(t)) h(t) y_i \mathrm{x}_i^\top} y_i \mathrm{x}_i.$$

For the last equation, we can require $g'(t) = e^{-f(g(t)}$. Approximating it with $g'(t) \approx e^{-f(g(t)) - \ln f'(g(t))}$ gives us a closed from solution $g(t) = f^{-1}(\ln t + C)$.

A.2 The LIGHT function

Let us substitute the default loss function $\ell(t)$ with the function that describes the behavior of the LIGHT neuron given in Section 1.1. It inherits the principles of population dynamics (Gray & Gray, 2017). In contrast to the family of the functions $\ell(t)$ thoroughly explored by Soudry et al. (2018); Nacson et al. (2019), the LIGHT function is non-monotonic and exhibits more complex behavior: it grows at a rate $r > 0$ according to the logistic law. After time $T \geq 0$, it declines at a rate $E \geq 0$. The y-intercept at the time instant T is equal to $0 < N_T < 1$. This behavior can be described as:

$$\ell^{\mathrm{LIGHT}}(t) = a e^{b e^{-r(t-d)}}, \tag{3}$$

for which $\ell'^{\mathrm{LIGHT}}(t) = -abre^{-f(t)}$, where $f(t) = r(t-d) - be^{-r(t-d)}$, $a = e^{\varepsilon}$, $\varepsilon = \frac{E}{r}$, $b = \ln N_T - \varepsilon$, $d = T$.

According to the reasoning presented in Section A.1, estimating the inverse function of $f(t)$ gives the convergence rate:

$$g^{\mathrm{LIGHT}}(z) = d + \big(\mathrm{W}_0(be^{-z}) + z\big)/r, \quad z > 0, \tag{4}$$

where $\mathrm{W}_0(z)$ is the principal branch of the Lambert function. Nacson et al. (2019); Soudry et al. (2018) showed that for any strict monotone loss $\ell(t)$, given in Section A.1, under certain conditions, $g(t) = \ln t$. With a variable substitute $z = \ln t$, the convergence rate $g(z) = z$ is further referred to as the default rate.

A.3 Convergence analysis

Let us first show that the LIGHT function changes the bounds of the default rate. Then, analyzing the LIGHT parameters, we confirm the validity of the laws given in Section 1.5.

Theorem. *For any $z > 0$, $b < 0$, moderate $r > 0$, and $d = 0$, the bounds of the convergence rate $g^{\mathrm{LIGHT}}(z)$ given by equation 4 are below and above the default convergence rate $g(z) = z$.*

Proof. Let us first analyze the parameters b, d, and r, which affect the convergence rate equation 4. We can see that $d > 0$ increases the absolute value of $g^{\mathrm{LIGHT}}(z)$. The parameter b depends on N_T

and the ratio E/r: $b = \ln N_T - E/r$ (see equation 3). As $0 < N_T < 1$, $\ln N_T < 0$. The smaller N_T is, the faster $|\ln N_T|$ increases. The ratio $E/r > 0$ grows up if $r \to 0$ (an infinitesimal value) or/and $E > r$. The parameter $b < 0$, but smaller N_T, r and larger E increase its absolute value.

Let us explore the bounds of $g^{\text{LIGHT}}(z)$. For $z > 0$, the equation $we^w = z$ has one positive solution $w = \mathrm{W}_0(z)$, which increases with z. If $z = e$, then $w = 1$. Thus, $w > 1$ if $z > e$. By taking logarithms of both sides, we get:

$$\begin{aligned} \ln w + w &= \ln z; \\ w &= \ln z - \ln w < \ln z. \end{aligned} \tag{5}$$

When $z > e$,

$$\begin{aligned} &1 < w < \ln x \\ &0 < \ln w < \ln \ln z. \end{aligned} \tag{6}$$

Substituting equation 6 into equation 5 yields:

$$\ln z - \ln \ln z < w < \ln z, \tag{7}$$

where the left side is positive for $z > 1$. Since $w = \mathrm{W}_0(z)$, we can write:

$$\ln z - \ln \ln z < \mathrm{W}_0(z) < \ln z, \tag{8}$$

Let us now modify the argument of $\mathrm{W}_0(z)$ with regard to $g^{\text{LIGHT}}(z)$:

$$\begin{aligned} \frac{b}{z} + z < \mathrm{W}_0(be^{-z}) + z &< \frac{b}{z} - \frac{b}{\ln z} + z; \\ \frac{b}{z} + z < \mathrm{W}_0(be^{-z}) + z &< b\frac{\ln z - z}{z \ln z} + z, \end{aligned}$$

where $\ln z - z < 0$ as $\ln z < z$ for all $z > 0$.

By definition, $b < 0$. Thus, for $z > e$:

$$\begin{aligned} b\frac{\ln z - z}{z \ln z} + z &> z; \\ \frac{b}{z} + z &< z \end{aligned}$$

As we see, the boundaries of $\mathrm{W}_0(be^{-z}) + z$ are below and above the default convergence rate z. □

Corollary. *The bounds of $g^{\text{LIGHT}}(z)$ move to the left when r is larger and to the right when d is larger and r is smaller.*

Proof. The validity of the corollary follows from equation 4. □

A.4 Validation of laws

From the convergence analysis, we can observe the behavior described in Section 1.4.

Increasing the convergence rate $g^{\text{LIGHT}}(z) \uparrow$ changes the parameters as follows:

$$r \uparrow \to r \downarrow; E \downarrow \to E \downarrow; T \uparrow \to T \uparrow; N_T \downarrow \to N_T \uparrow,$$

where ↑ denotes an increase, ↓ stands for a decrease, and → represents a change in direction with time. From this chain, we can conclude that the LIGHT population intends to keep its size close to the maximum with no growth and harvesting, which means with the maximum internal competition. This confirms the validity of **Law 1** given in Section 1.5.

Decreasing the convergence rate $g^{\text{LIGHT}}(z) \downarrow$ leads to the following changes:

$$r \downarrow \to r \uparrow; E \downarrow \to E \uparrow; T \downarrow \to T \downarrow; N_T \downarrow \to N_T \downarrow.$$

Here, we can see that the LIGHT population intends to minimize the size with higher growth and harvesting rates, which results in lower competition. This validates **Law 2**.