PLOS ONE

RESEARCH ARTICLE

BioGD: Bio-inspired robust gradient descent

Ilona Kulikovskikh[1,2,3]*, **Sergej Prokhorov**[1], **Tomislav Lipić**[3], **Tarzan Legović**[4], **Tomislav Šmuc**[3]

1 Department of Information Systems and Technologies, Samara National Research University, Samara, Russia, **2** Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia, **3** Division of Electronics, Ruđer Bošković Institute, Zagreb, Croatia, **4** Division of Marine and Environmental Research, Ruđer Bošković Institute, Zagreb, Croatia

* ilona@irb.hr

Citation: Kulikovskikh I, Prokhorov S, Lipić T, Legović T, Šmuc T (2019) BioGD: Bio-inspired robust gradient descent. PLoS ONE 14(7): e0219004. https://doi.org/10.1371/journal.pone.0219004

Editor: Sanda Martinčić-Ipšić, University of Rijeka, CROATIA

Received: February 1, 2019

Accepted: June 13, 2019

Published: July 5, 2019

Data Availability Statement: All relevant data are within the manuscript, Supporting Information files, and on Github: https://github.com/yukinoi/bio_gradient_descent/tree/master/biogd/data.

Funding: This work was supported by the Ministry of Science and Higher Education of Russian Federation (Russian Federation President grant No. MK-6218.2018.9) to IK, the Russian Foundation for Basic Research (grant No. 18-37-00219) to IK, and the Ministry of Education and Science of the Russian Federation (grant No. 074-U01) to IK, and SP. IK, ToL and TS acknowledge the support by the

Abstract

Recent research in machine learning pointed to the core problem of state-of-the-art models which impedes their widespread adoption in different domains. The models' inability to differentiate between noise and subtle, yet significant variation in data leads to their vulnerability to adversarial perturbations that cause wrong predictions with high confidence. The study is aimed at identifying whether the algorithms inspired by biological evolution may achieve better results in cases where brittle robustness properties are highly sensitive to the slight noise. To answer this question, we introduce the new robust gradient descent inspired by the stability and adaptability of biological systems to unknown and changing environments. The proposed optimization technique involves an open-ended adaptation process with regard to two hyperparameters inherited from the generalized Verhulst population growth equation. The hyperparameters increase robustness to adversarial noise by penalizing the degree to which hardly visible changes in gradients impact prediction. The empirical evidence on synthetic and experimental datasets confirmed the viability of the bio-inspired gradient descent and suggested promising directions for future research. The code used for computational experiments is provided in a repository at https://github.com/yukinoi/bio_gradient_descent.

Introduction

Modern machine learning algorithms can successfully tackle tough and complicated problems by taking patterns buried inside datasets to build a model with remarkable predictive capabilities [1]. However, the machine learning models fueling innovations in a variety of applications from large-scale genomic sequencing and medicine [2–6] to automated driving [7] and robotics [8] still pose serious challenges that make it difficult to fully trust and adopt them [7, 9, 10]. A lack of intelligence in these models leads to an inability to robustly differentiate between noise and subtle, but significant variation in data. If the noise in data includes intentionally small carefully crafted perturbations, which are used to generate so-called adversarial examples, the model may become vulnerable and misclassify them with high confidence [11–17]. It

Centre of Excellence project "DATACROSS," co-financed by the Croatian Government and the European Union through the European Regional Development Fund—the Competitiveness and Cohesion Operational Programme (KK.01.1.1.01.0009). The URLs of the funders are https://minobrnauki.gov.ru, http://www.rfbr.ru/rffi/eng and https://ec.europa.eu/regional_policy/en/funding/erdf/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

expectedly sets up the psychological roadblocks [7] to the widespread adoption of machine learning models in different domains [3, 7, 11, 12, 17].

Addressing the question of vulnerability, the researchers created various attacking strategies [18–23] to fool the model with adversarial examples as well as defenses [24–29] to resist them, but that has not solved the problem completely. The common approach to creating attacks is taking gradients with regard to their inputs, since gradients provide local linear approximations of models behavior [12]. Goodfellow et al. [11] explored the connection between models vulnerability and linear nature [30, 31] which allows introducing infinitesimal changes in the gradients. According to the authors' reasoning, the limited precision of features leads to the identical interpretation on the input gradients and adversarial gradients. With well-separated classes, the models assign the same class for both the input and adversarial gradients as long as the magnitude of perturbations is less than the magnitude of the input gradients which are predefined by the precision of features. For this reason, it seems rational to penalize the large input gradients [12, 13] which are more likely to be utilized for generating adversarial examples. The gradient regularization methods employ the idea of a double backpropagation method first introduced by Drucker and Le Cun [32]: training neural networks by minimizing not only network "energy" but also the rate of energy changes with regard to the input features. Adopting such gradient regularization on a training dataset may increase robustness to adversarial perturbations embedded in an unseen dataset as much as adversarial training [12].

What is not specifically undertaken in the studies mentioned above is whether the models inspired by biological evolution [33–42] may result in better robustness to adversarial perturbations. This is the research question raised in this study. The fundamental aspects of biological intelligence, such as self-healing, evolution, and learning make biological organisms successful to survive in unknown and changing environments [33]. The stability and adaptability of biological systems strengthen the motivation for replicating the mechanisms of natural evolution in an attempt to create the models with characteristics comparable to those of biological systems.

This paper seeks to refine the discourse on robustness to adversarial noise with the bio-inspired gradient descent based on the generalized Verhulst population growth equation [43, 44]. The hyperparameters of the Verhulst equation are used to regularize gradients or, to put it more specifically, to penalize the degree to which imperceptible changes in gradients may influence prediction results. We refer to them as *the lower and upper levels of visibility* as they limit the gradients to be near zero so that any small magnitude perturbation hidden in the gradients is "visible" and, then, has no influence on prediction results.

Materials and methods

Gradient descent

Similarly to the learning setting presented by Soudry et al. [45] we consider a dataset $\{\mathrm{x}_i, y_i\}_{i=1}^m$ with $\mathrm{x}_i \in \mathbb{R}^n$, $y_i \in \{0, 1\}$ and minimize an empirical loss function

$$\mathscr{L}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \ell(\boldsymbol{\theta}^T \mathrm{x}_i),$$

with a weight vector $\boldsymbol{\theta} \in \mathbb{R}^n$. We are interested in linearly separable problems with a smooth monotone strictly decreasing and non-negative loss function (see Assumption 1 and 2). For clarity, the learning setting is illustrated in Fig 1.

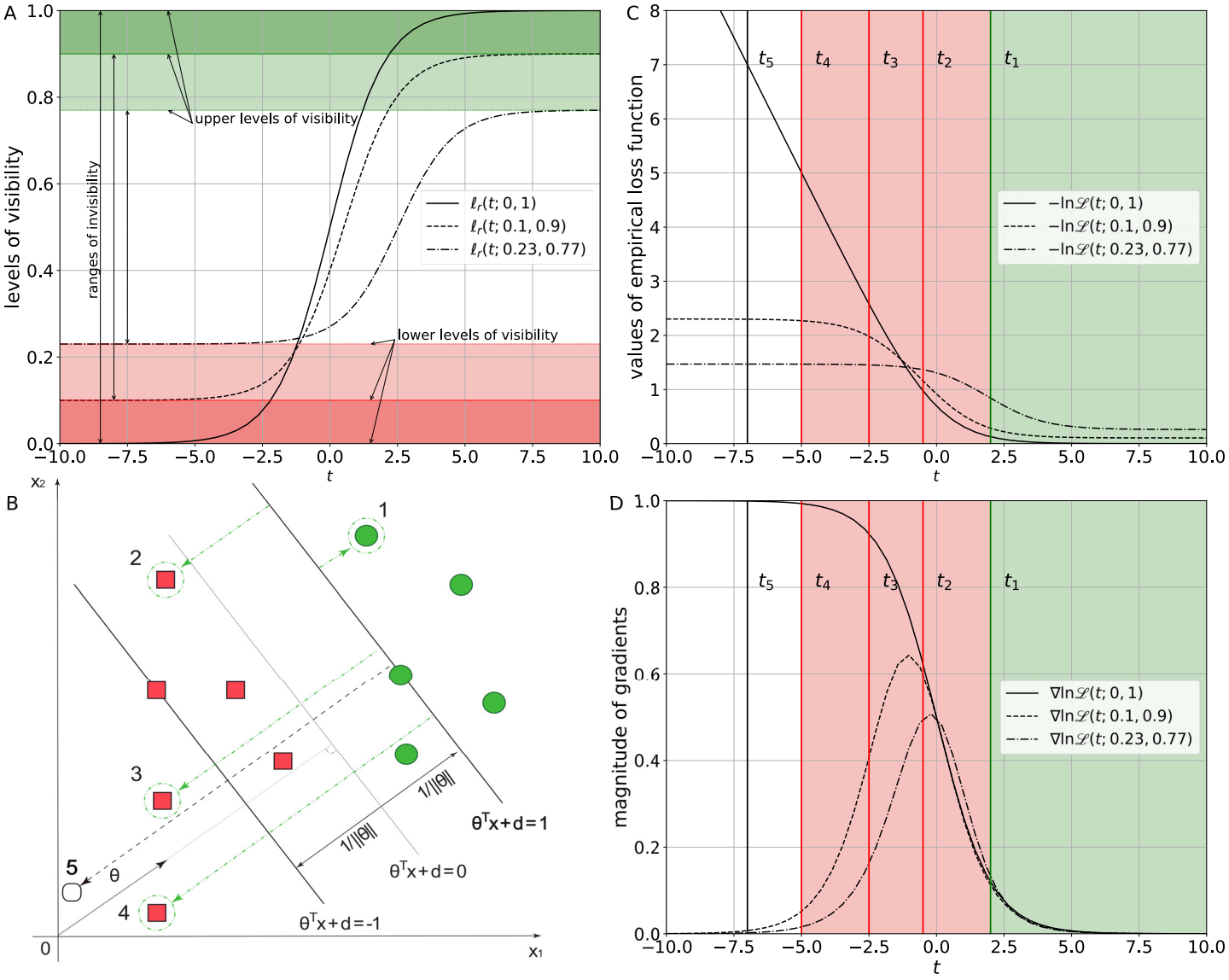

Fig 1. The influence of a difference in approaching the lower and the upper asymptote on the magnitude of gradients. (A) The definitions of levels of visibility and ranges of invisibility based on the generalized logistic loss function $\ell_r(t; a, b)$ subject to the different pairs $(a, b) = \{(0, 1), (0.1, 0.9), (0.23, 0.77)\}$ and $r = 1$. (B) A decision boundary with reliable (1), noisy (2,3,4), and adversarial (5) instances. (C) The empirical generalized logistic loss function $-\ln \mathcal{L}(t; a, b)$ with the levels of visibility for reliable t_1, noisy t_2, t_3, t_4, and adversarial t_5 instances subject to the different pairs $(a, b) = \{(0, 1), (0.1, 0.9), (0.23, 0.77)\}$ and $r = 1$. (D) The magnitude of the gradients of the empirical generalized logistic loss function $\nabla\ln \mathcal{L}(t; a, b)$ subject to the different pairs $(a, b) = \{(0, 1), (0.1, 0.9), (0.23, 0.77)\}$ and $r = 1$.

https://doi.org/10.1371/journal.pone.0219004.g001

Assumption 1. *The dataset is linearly separable*: $\exists\, \boldsymbol{\theta}^*$ *such that* $\forall i$: $\boldsymbol{\theta}^{*T} \mathrm{x}_i > 0$.

Assumption 2. $\forall t \in \mathbb{R}$: $\ell(t)$ *is a differentiable, monotonically decreasing function bounded from below*: $\ell(t)>0$, $\ell'(t)<0$, $\lim_{t\to\infty} \ell(t) = \lim_{t\to\infty} \ell'(t) = 0$.

The solution to the problem $\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \mathscr{L}(\boldsymbol{\theta})$ can be found using the l^{th} iteration of the gradient descent *GD* updates with a learning rate η:

$$\boldsymbol{\theta}_{l+1} = \boldsymbol{\theta}_l - \eta\nabla\mathscr{L}(\boldsymbol{\theta}_l) = \boldsymbol{\theta}_l - \eta\sum_{i=1}^{m}\ell'(\boldsymbol{\theta}_l^{\mathrm{T}}\mathrm{x}_i)\mathrm{x}_i. \tag{1}$$

In Eq (1) it is assumed that $\forall i \in \{1, \ldots, m\}$: $y_i = 1$, $\|x_i\|<1$.

The function $\ell(t)$ describes a common classification loss function, including the exponential and logistic loss. The generalized Verhulst model allows us to introduce the generalized logistic loss function. In contrast to the simple loss function, the generalized loss function involves a lower and an upper asymptote that allow additional regulating the magnitude of input gradients.

Generalized Verhulst growth model

The population growth $P(t)$ in an infinite environment can be described with a linear ordinary differential equation [44, 46, 47]:

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = rP(t), \tag{2}$$

where $r \equiv b - d > 0$ is the *per capita* rate of population growth, b and d are respective *per capita* rates of birth and death. The solution to this equation is

$$P(t) = P_0 \exp(rt), \tag{3}$$

where P_0 is the initial population.

Eq (2) introduces the Malthus model that describes an exponential growth of the population [47]. The Malthus model seems sufficient to describe the population in the initial stage of growth while the population is still small. However, it fails on a longer time interval since it leads to an infinite growth $\lim_{t\to\infty} P(t) = \infty$ as it does not consider the scarcity of resources.

To overcome this disadvantage, Verhulst assumed that Eq (2) should include the correction term [43, 46]. This term approaches 1 when $P(t)$ tends to zero and decreases linearly as $P(t)$ increases. Thus, the Malthus equation takes the form of the Verhulst equation that includes the correction term proportional to $-P(t)^2/K$:

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = rP(t)\left(1 - \frac{P(t)}{K}\right), \tag{4}$$

where K is the carrying capacity that represents the maximum size of population which can be supported by the environment. As the population size approaches the carrying capacity $\lim_{t\to\infty} P(t) = K$, the population stops growing.

The solution to Eq (4) introduces the logistic growth model:

$$P(t) = \frac{K}{1 - \left(1 - \frac{K}{P_0}\right)\exp(-rt)}. \tag{5}$$

The extension to Eq (4) that also considers a critical population size can be rewritten as

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = r(P(t) - A)\left(1 - \left(\frac{P(t) - A}{K - A}\right)\right); \tag{6}$$

$$P(t) = A + \frac{K - A}{1 - \left(1 - \left(\frac{K-A}{P_0+A}\right)\right)\exp(-rt)}, \tag{7}$$

where K is the upper asymptote or the population carrying capacity; A is the lower asymptote or the population minimum size.

The asymptote A indicates critical population thresholds $0 \leq A < K$ below which a population crashes to extinction. It serves as a substitute for the Allee effects [48] which are broadly defined as a decline in individual fitness at low population size or density.

Bio-inspired gradient descent

We introduce the generalized logistic loss function $\ell_r(t;a, b)$ with regard to the generalized Verhulst growth model Eq (6) and its solution Eq (7):

$$\ell'_r(t;a,b) = r(\ell_r(t;a,b) - a)\left(1 - \frac{\ell_r(t;a,b) - a}{b - a}\right) \tag{8}$$

$$\forall t \in \mathbb{R} : \ell'_r(t;a,b) < 0, \lim_{t\to\infty} \ell'_r(t;a,b) = \lim_{t\to-\infty} \ell'_r(t;a,b) = 0,$$

$$\ell_r(t;a,b) = a + \frac{b-a}{\left(1 - \left(1 - \left(\frac{b-a}{P_0 + a}\right)\right)\exp(-rt)\right)}, \tag{9}$$

so that $\forall t \in \mathbb{R} : \ell_r(t;a,b) > 0, \lim_{t\to\infty} \ell_r(t;a,b) = b - a, \lim_{t\to-\infty} \ell_r(t;a,b) = a$, where the lower asymptote $a \equiv A$, the upper asymptote $b \equiv K$, $0 \le a < b \le 1$ and $P_0 + a = \frac{b-a}{2}$, from which $P_0 = \frac{b-3a}{2}$.

Eqs (8) and (9) can be presented as

$$\ell'_r(t;a,b) = \frac{(b-a)r\exp(-rt)}{1 + \exp(-rt)} = (b-a)r\ell'(t), \tag{10}$$

$$\ell_r(t;a,b) = a + \frac{b-a}{1 + \exp(-rt)}, \tag{11}$$

where $\ell'(t) = \ell(t)(1 - \ell(t))$ is the derivative of the simple logistic loss function $\ell(t) = \ell_1(t;0, 1)$.

Using Eq (10) to determine the gradient in Eq (1) allows us to present the bio-inspired gradient descent *BioGD* updates:

$$\boldsymbol{\theta}_{l+1} = \boldsymbol{\theta}_l - \eta\sum_{i=1}^{m}\ell'_r(\boldsymbol{\theta}_l^{\mathrm{T}}\mathrm{x}_i;a,b)\mathrm{x}_i = \boldsymbol{\theta}_l - \eta_r\sum_{i=1}^{m}\ell'(\boldsymbol{\theta}_l^{\mathrm{T}}\mathrm{x}_i)\mathrm{x}_i, \tag{12}$$

where $\eta_r = (b-a)r\eta$.

For the sake of completeness, we introduce the empirical generalized logistic loss function and its gradient as follows:

$$\ln\mathscr{L}(\boldsymbol{\theta};a,b) = -\sum_{i=1}^{m} y_i \ln\ell_r(\boldsymbol{\theta}^T\mathrm{x}_i;a,b) + (1-y_i)\ln(1 - \ell_r(\boldsymbol{\theta}^T\mathrm{x}_i;a,b)),$$

$$\nabla\ln\mathscr{L}(\boldsymbol{\theta};a,b) = -\sum_{i=1}^{m}(y_i - \ell_r(\boldsymbol{\theta}^{\mathrm{T}}\mathrm{x}_i;a,b))\frac{\ell'_r(\boldsymbol{\theta}^{\mathrm{T}}\mathrm{x}_i;a,b)}{(1 - \ell_r(\boldsymbol{\theta}^{\mathrm{T}}\mathrm{x}_i;a,b))\ell_r(\boldsymbol{\theta}^{\mathrm{T}}\mathrm{x}_i;a,b)}\mathrm{x}_i, \forall y_i \in \{0,1\}.$$

Then, taking into account our assumption that $y_i = 1$, Eq (12) can be given as:

$$\boldsymbol{\theta}_{l+1} = \boldsymbol{\theta}_l - \eta\sum_{i=1}^{m}\frac{\ell'_r(\boldsymbol{\theta}_l^{\mathrm{T}}\mathrm{x}_i;a,b)}{\ell_r(\boldsymbol{\theta}_l^{\mathrm{T}}\mathrm{x}_i;a,b)}\mathrm{x}_i = \boldsymbol{\theta}_l - \eta_r\sum_{i=1}^{m}\frac{\ell'(\boldsymbol{\theta}_l^{\mathrm{T}}\mathrm{x}_i)}{\ell_r(\boldsymbol{\theta}_l^{\mathrm{T}}\mathrm{x}_i;a,b)}\mathrm{x}_i. \tag{13}$$

The *BioGD*'s gradients in Eq (13) involve the derivative of the logistic loss function $\ell'(\boldsymbol{\theta}_l^{\mathrm{T}}\mathrm{x}_i)$ which is directly used to define the Hessian matrix of the simple logistic loss function in a second-order gradient method. This means that we can regularize the first-order gradients by partly employing the beneficial properties of the second-order method.

By definition, $b - a \leq 1$ that implies $P_0 \leq \frac{1}{2} - a$. Let $a^*(b)$ be the solution to $f(a^*(b);b) = 0$, where $f(a; b) = \frac{b-3a}{2}$. Then, $P_0 < a$ if $a > a^*(b)$. This condition guarantees that the lower asymptote of the generalized logistic loss function $\ell_r(t; a, b)$ is approached much more gradually than the upper asymptote. In other words, the lower plateaus of the generalized logistic loss function $\ell_r(t; a, b)$ with the nontrivial values $(a, b) = \{(0.1, 0.9), (0.23, 0.77)\}$ are broader than its upper plateaus in comparison with the simple logistic loss function $\ell(t)$ (see Fig 1A).

The solution $a^*(b)$ influences the rate of the loss function growth $\ell_r(t; a, b)$ if $t < 0$. The value of the lower asymptote has impact on the robustness of gradient descent if $\forall i \in \{1, \ldots, m\}$: $t = \boldsymbol{\theta}^{*T} \mathrm{x}_i < 0$, i.e. in the presence of noisy and adversarial labels. Fig 1B depicts a decision boundary with reliable **1**, noisy **2**, **3**, **4**, and adversarial **5** instances. The circles denoted by **1** are true positive instances. The squares present true negative instances. The instances indexed by **2**, **3**, **4**, but annotated as a positive class, are noisy labels. An adversarial instance denoted by **5** is moved away from its legitimate class in the direction orthogonal to $\boldsymbol{\theta}$ and, as a result, is "invisible" and misclassified by the model.

The labels marked in Fig 1B specify the reliable ($t \geq 0$) and noisy ($t \leq 0$) regions for the empirical logistic loss function and its gradient in Fig 1C and 1D. As it can be seen, the difference in approaching the lower and upper levels of visibility produces a positive effect on the loss function by dropping noisy t_1, t_2, t_3, t_4 and adversarial t_5 instances. In addition, it limits the gradients to be near zero that allows any small magnitude perturbations embedded in the gradients to be "visible" and make sure that they have no influence on classification results.

Therefore, the existing difference in approaching the two asymptotes allows us to expose small magnitude perturbations below some lower and upper asymptotes, i.e. the lower and upper levels of visibility, and diminish its negative effect on results. In terms of population dynamics, the range of invisibility between these levels, i.e. the minimum and the maximum population size may be interpreted as the population persistence in spite of very limited food supply or space.

BioGD

BioGD represents an implementation of bio-inspired gradient descent (see Algorithm). The algorithm involves the optimization of hyperparameters a, b and r with a grid search. A grid with sufficient granularity to optimizing hyper-parameters can be easily implemented but requires considerable computational costs if there is a need for smaller grid step sizes. In addition, the hyperparameters need to be optimized with constraints imposed by their bio-inspired interpretation in terms of the generalized Verhulst equation. An inappropriate choice of grid steps taken by the constrained optimization problem may lead to a lack of convergence. As an alternative to grid search, a random search may be taken into consideration [49]. This technique combines the hyperparameters randomly and demonstrates comparatively better results. However, as the method is entirely random, it leads to high variance. Bayesian optimization has lower variance as it searches the hyperparameters relying on probabilistic inference [50]. This virtue may, however, induce bias in the estimates of hyperparameters in the presence of outliers. Consequently, the choice of method depends on specific requirements and hyperparameter space complexity related to the problem to be solved. In this paper, we focus on the problem of robustness to outliers, in particular, adversarial noise. For this reason, we opt for a grid search to provide more reliable results by the exclusion of unnecessary variance in the estimates of hyperparameters.

Algorithm Bio-inspired gradient descent

```
procedure BioGD (x, y, η, n)
  Initialize θ_0;
  Initialize a ∈ [a_min, a_max], b ∈ [b_min, b_max], r ∈ [r_min, r_max];
```

```
Initialize a grid of n points in the space a × b × r;
Split (x, y) into train (x, y)_T and cross-validation (x, y)_CV subsets;
l ← 0;
repeat
  θ_{l+1} ← θ_l − (b − a) r η ∇_(θ) ℒ(θ_l, (x, y)_T);
  l ← l + 1;
until converge
(a, b, r) ← GridSearch(ℒ(θ_{l+1}, (x, y)_CV), a, b, r);
return θ_{l+1}, (a, b, r)
```

Computational experiments

We analyzed the impact of the hyperparameters on the robustness properties empirically in order to support the results of theoretical outcomes. According to theoretical results, the hyperparameters a and b: 1) directly deal with the magnitude of gradients; 2) are adaptive to the newly updated data; 3) allow us to improve robustness to noise by penalizing the degree to which hardly visible changes in gradients impact prediction. Therefore, the empirical evidence is based on the comparison between the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with a_{opt}, b_{opt}. In addition, we varied incrementally a number of instances m to explore the adaptivity of the hyperparameters a_{opt}, b_{opt} to the newly updated data. Finally, we justified the improvement in robustness and performance of the generalized model over the simple model on the synthetic linearly separable dataset with different proportions of noisy labels. Then, we assessed the viability of the generalized logistic loss with the hyperparameters in the more realistic setting on experimental datasets. The chosen datasets are normally not linearly separable but are indicative of the behavior of the hyperparameters and their influence on prediction results.

BioGD suggests a predefined value for the rate of population growth for computational experiments on synthetic and experimental datasets. There are two reasons for choosing this value. First, we intend to analyze the direct influence of the hyperparameters a and b on the magnitude of gradients. Second, we explore the benefits of the generalized logistic loss with some optimal hyperparameters a_{opt} and b_{opt} over the simple logistic loss with a = 0 and b = 1. The simple logistic loss lacks the parameter r by definition. As a consequence, it is reasonable to choose $r = 1$.

While the hyperparameters a and b address the issue of robustness, the growth rate r is aimed at accelerating the rate of convergence adaptively. Consequently, the hyperparameter r does not play a crucial role in improving the results on the synthetic and experimental datasets since the results of computational experiments are given from the perspective of a discriminative model such as logistic regression for binary classification. This normally implies processing smaller datasets. Even if the detailed analysis of the convergence rate issue is beyond the scope of this study, we still took the parameter r into consideration to show the usefulness and applicability while applying the proposed technique to a more complicated model on a large dataset for multiclass image classification.

Synthetic datasets

Design of experiments. We created an n-dimensional dataset so that each dimension's mean μ_j is sampled from a Gaussian distribution $\mathcal{N}(0, 1)$ while each dimension's standard deviation σ_j is generated according to a distribution $\mathcal{N}(1, 1)$. The feature space of n-dimensional instances is sampled from the distribution $\mathcal{N}(\mu_j, \sigma_j)$. The instances in the dataset were labelled by a randomly chosen hyperplane, so that the generated dataset is linearly separable. A

similar design of experiments is given by [51]. For the originally generated dataset, the portion of mislabelled examples is $p = 0.0$. We constructed the noisy versions of this dataset by flipping the labels for the randomly selected proportion of data instances $p = \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. A more detailed description of this procedure is presented by [52] The experiments were carried out by changing the number of instances incrementally $m_k \in [20, 100]$ and setting the number of features $n = 10$.

The datasets were divided into the training subset and the validation subset using 5-fold cross-validation. We evaluated the cross-validation estimates of the hyperparameters a_{opt} and b_{opt} with BioGD refined over a grid of 20 points in the hyperparameter space $a \times b$, where $a \in [0, 0.15]$ and $b \in [0.85, 1]$. In addition, we computed the mean and standard deviation (sd) of empirical logistic loss $\ln \mathscr{L}(\boldsymbol{\theta}; x^{m_k}, a_{opt}, b_{opt})$ over $N = 200$ experiments to guarantee the reliable estimates. We used the default value for the learning rate $\eta = 1.5e\text{-}8$ to run BioGD.

Results. Fig 2A, 2C and 2E present the estimates of mean and sd of empirical loss for the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} with regard to $m_k = [20, 100]$ and $p = \{0.0, 0.15, 0.3\}$. As it can be seen, the generalized logistic loss with the asymptotes a_{opt}, b_{opt} brings the benefits to both mean and sd of the empirical loss. Moreover, increasing the portion of noisy labels up to $p = 0.3$ (see Fig 2E) allows us to have distinct advantage of introducing the loss with optimal levels of visibility a_{opt}, b_{opt} over the simple loss with a = 0, b = 1 on linearly separable dataset $p = 0.0$ (see Fig 2A). The impact of levels of visibility on the estimates may be seen in Fig 2B, 2D and 2F. We have linearly separable datasets in Fig 2B. The lower level of visibility does not come into play here as $a_{opt} = 0$, but the upper level of visibility handles premature convergence of the empirical loss. On the other hand, different portions of noisy labels were presented in Fig 2D and 2F, where

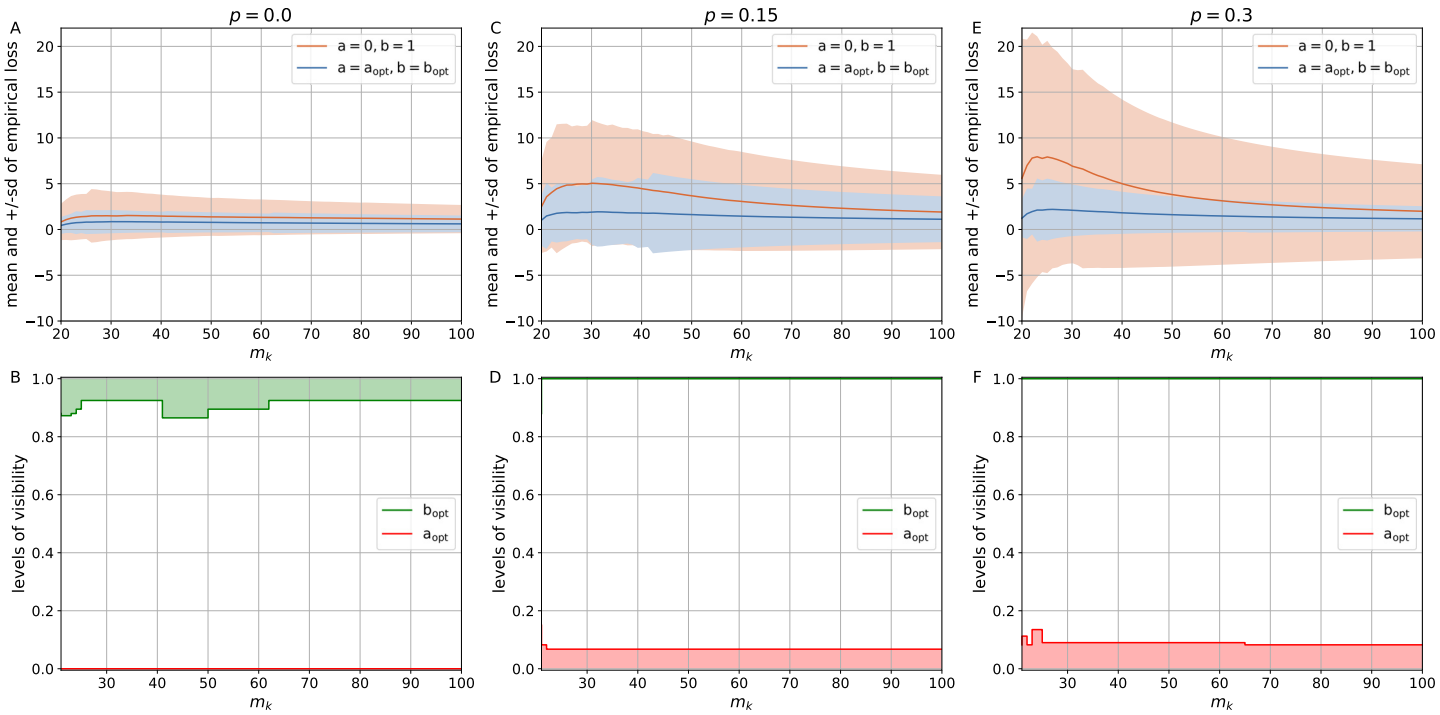

Fig 2. The estimates for the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to $m_k \in [20, 100]$ and $p = \{0.0, 0.15, 0.3\}$. (A) The mean and +/-sd of empirical loss subject to $p = 0.0$. (B) The levels of visibility a_{opt}, b_{opt} subject to $p = 0.0$. (C) The mean and +/-sd of empirical loss subject to $p = 0.15$. (D) The levels of visibility a_{opt}, b_{opt} subject to $p = 0.15$. (E) The mean and +/-sd of empirical loss subject to $p = 0.3$. (F) The levels of visibility a_{opt}, b_{opt} subject to $p = 0.3$.

https://doi.org/10.1371/journal.pone.0219004.g002

we can observe how the lower level of visibility minimizes both the mean and sd of the empirical loss. If the lower level of visibility a_{opt} results in larger values, it produces a more dramatic effect on the estimates of empirical loss. This is expected by definition as the lower level allows us to directly regulate the magnitude of gradients (see Fig 1D). The results of experiments for different pairs $p = \{0.05, 0.1\}$ and $p = \{0.2, 0.25\}$ are given in S1 and S2 Figs, respectively.

To confirm the viability of introducing the levels of visibility, we computed the estimates summed over the interval $m_k = [20, 100]$ for the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} with regard to different proportions $p = \{0.0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ (see Fig 3). As we can see, there is a more noticable improvement in both the mean (see Fig 3A) and sd (see Fig 3B) of the empirical loss for larger portion of noisy labels that fully comply with the theoretical reasoning behind the definitions of visibility levels presented in Fig 1. The estimates of mean and +/-sd of empirical loss for $p = \{0.0, 0.3\}$ are given in S1 and S2 Files, respectively.

We used the Wilcoxon signed-rank test [53] to confirm the statistical significance of the differences in mean against sd of the empirical loss between the generalized logistic loss and the simple logistic loss without assuming them to follow the normal distribution. This test is usually considered an alternative to the paired Student's t-test if the population can not be assumed to be normally distributed. The null hypothesis is that there is no improvement in the mean and sd of the empirical loss for the generalized logistic loss with the optimal levels of visibility over the empirical loss for the simple logistic loss. Since the p-value turns out to be 0.01563, i.e. less than the 0.05 significance level, we rejected the null hypothesis. Consequently, the difference between the mean and sd of the empirical loss for the generalized logistic loss and the simple loss is statistically significant. In addition, we applied a one-sided Wilcoxon test with Bonferroni correction [54] for each p. The tests resulted in p-value<0.001 which means that for each value of p we have a sufficient reason to reject the null hypothesis.

Experimental datasets

Design of experiments. We chose 11 experimental datasets, which are freely available from UCI Machine Learning repository, for validating BioGD. The information given in

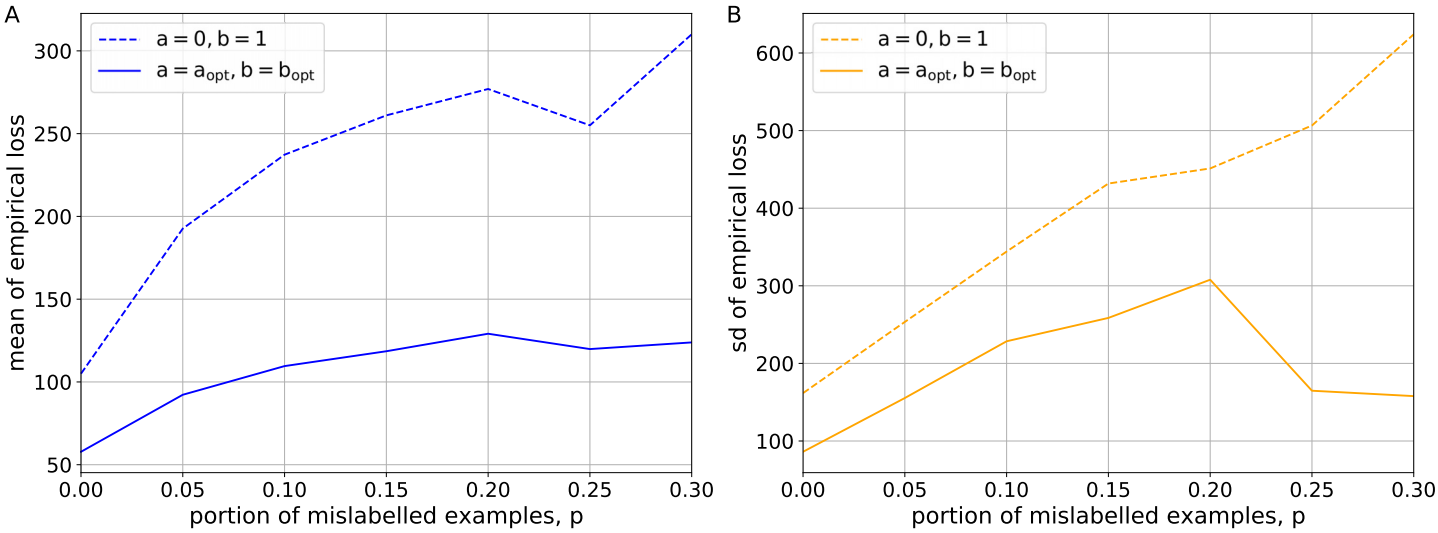

Fig 3. The estimates for the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to the proportions of noisy labels $p = \{0.0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. (A) The mean of empirical loss summed over the interval $m_k = [20, 100]$. (B) The sd of empirical loss summed over the interval $m_k = [20, 100]$.

https://doi.org/10.1371/journal.pone.0219004.g003

Table 1 includes the number of instances m, the number of features n, the class distribution $|y_i = 0|$ and $|y_i = 1|$. Incrementally varying a number of instances $m_k \in [20, 100]$ randomly taken from the original dataset with number of instances m, we explored the adaptivity of the hyperparameters a_{opt} and b_{opt} to the newly updated data. By analogy with the empirical setting for the synthetic datasets, we implemented BioGD over a grid of 20 points in the hyperparameter space $a \times b$, where $a \in [0, 0.15]$ and $b \in [0.85, 1]$, to make the cross-validation estimates of the empirical loss and the asymptotes. The default learning rate η = 1.5e-8 was used.

Results. According to the results of computational experiments, the used datasets can be divided into the groups with marked, little and no improvement in the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} over the simple logistic loss with a = 0, b = 1. Figs 4 and 5 demonstrate the estimates of mean and sd of the empirical loss over $N = 200$ experiments with regard to $m_k = [20, 100]$ for *vertebral*, *liver*, *pima*, and *heart*. We can observe that the generalized logistic loss with the asymptotes a_{opt}, b_{opt} brings the benefits to both mean and sd of the empirical loss. Moreover, the loss function converges with an increasing number of examples m_k. As we have seen in case of the synthetic dataset, the variations of the upper level of visibility b_{opt} can be attributed to premature convergence of the empirical loss function, while the lower level of visibility is sensitive to the outliers that results in its divergence. Moreover, the larger values for either the lower or upper level of visibility are indicative of the domination of the above-mentioned issues.

However, for *banknote*, *breast-win*, and *climate* datasets, we obtained little improvement as both levels of visibility influence the results or their impact is negligible (see S3 Fig). This pinpoints a lack of salient presence of one of the above-mentioned issues. As a result, the empirical loss function may not converge with an increasing number of examples m_k. The levels of visibility may not produce any improvement (see S4 Fig). For *blood* dataset the curves of empirical loss function for the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} and the simple logistic loss with a = 0, b = 1 have been overlapped. For *parkinsons* dataset, the empirical loss functions have little advantage over the generalized loss. Both almost overlap. In S5 Fig the empirical loss functions does not converge for the simple logistic loss with a = 0, b = 1 at all. The corresponding curves have been set as zero-valued. But the optimal levels of visibility a_{opt}, b_{opt} still allow us to estimate the mean and sd of the generalized loss.

Fig 6 depicts the differences in mean against sd of the empirical loss between the simple logistic loss and the generalized logistic loss for all the datasets summed over the interval $m_k \in [20, 100]$. The circles, radius of which is proportional to the number of features n (see Table 1), are colored according to the rate of improvement on the mean and sd of empirical loss. As the

Table 1. A brief description of the experimental datasets.

Dataset	$m = \|y_i\|$	$m = \|y_i = 0\|$	$m = \|y_i = 1\|$	n
Blood Transfusion Service Center (blood)	748	570	178	4
Banknote Authentication (banknote)	1372	762	610	4
Vertebral Column (vertebral)	309	100	209	6
Liver Disorder (liver)	345	145	200	6
Pima Indians Diabetes (pima)	768	500	268	8
Breast Cancer Wisconsin (breast-win)	683	444	239	9
Heart Statlog (heart)	270	150	120	13
Climate Model Simulation Crashes (climate)	540	46	494	17
Parkinsons (parkinsons)	195	48	147	22
Chronic Kidney Disease (chronic-kidney)	157	114	43	24
Cervical Cancer (cervical)	668	623	45	33

https://doi.org/10.1371/journal.pone.0219004.t001

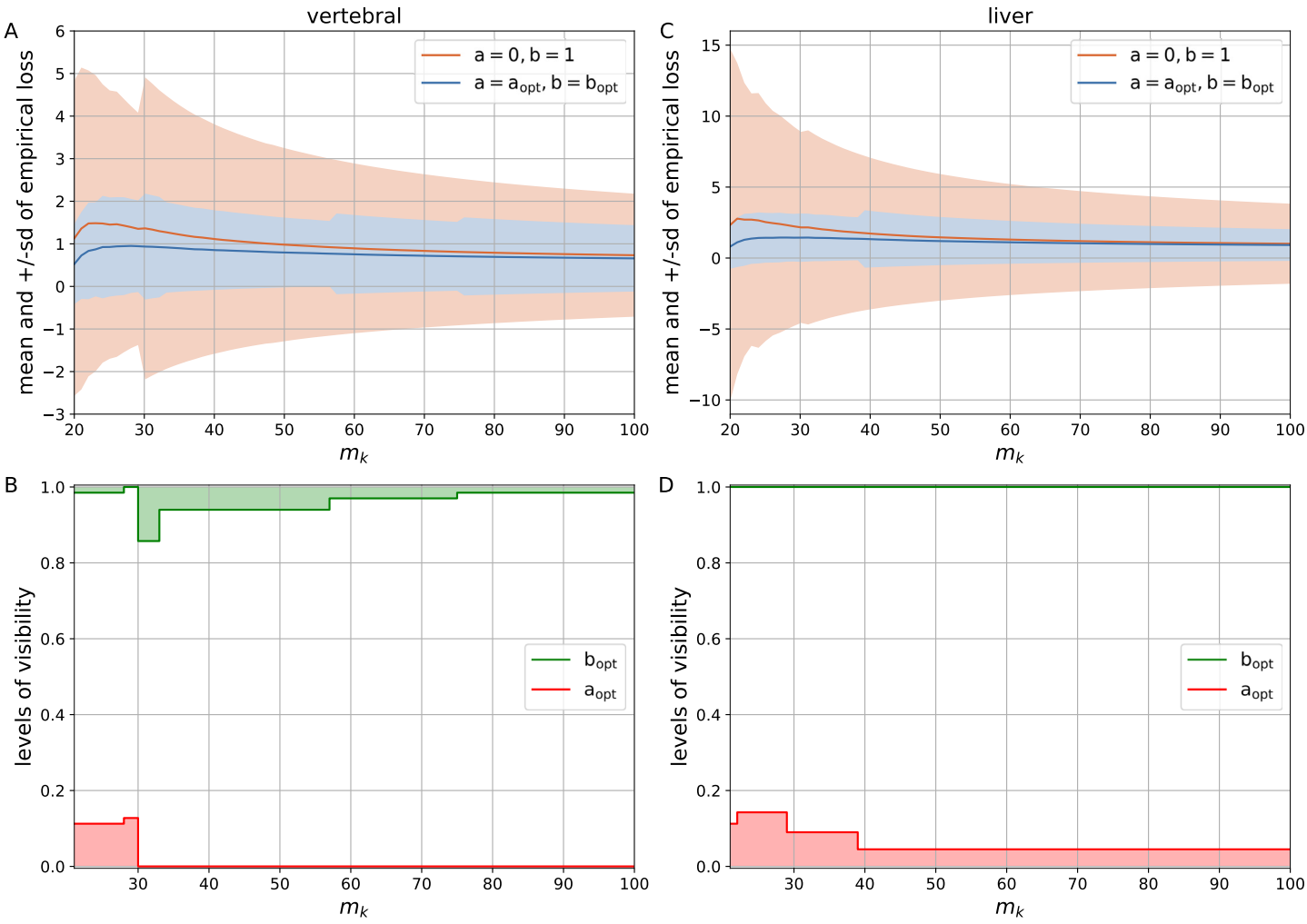

Fig 4. The estimates for the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to $m_k \in [20, 100]$. (A) The mean and +/-sd of empirical loss for the *vertebral* dataset. (B) The levels of visibility a_{opt}, b_{opt} for the *vertebral* dataset. (C) The mean and +/-sd of empirical loss for the *liver* dataset. (D) The levels of visibility a_{opt}, b_{opt} for the *liver* dataset.

https://doi.org/10.1371/journal.pone.0219004.g004

empirical loss function does not converge for *chronic-kidney* and *cervical* datasets in case of a = 0, b = 1, the illustrated differences do not include these datasets. The estimates of mean and sd for the simple loss and the generalized loss separately are given in S6 Fig.

By analogy with the synthetic datasets, we used the Wilcoxon signed-rank test to confirm the statistical significance of the differences in mean against sd of the empirical loss between the simple logistic loss and the generalized logistic loss. Since the p-value is equal to 0.01427, i.e. less than the 0.05 significance level, we rejected the null hypothesis. The difference between the mean and sd of the empirical loss for the generalized logistic loss and the simple loss is statistically significant.

Real-world application

Design of experiments. As an example of the applicability of the proposed optimization technique, we applied it to an Adam-based [55] neural network with two hidden layers, each of them containing 10 neurons. Each layer was activated by an advanced sigmoid function. The activation function of the neural network was built with the generalized logistic loss that

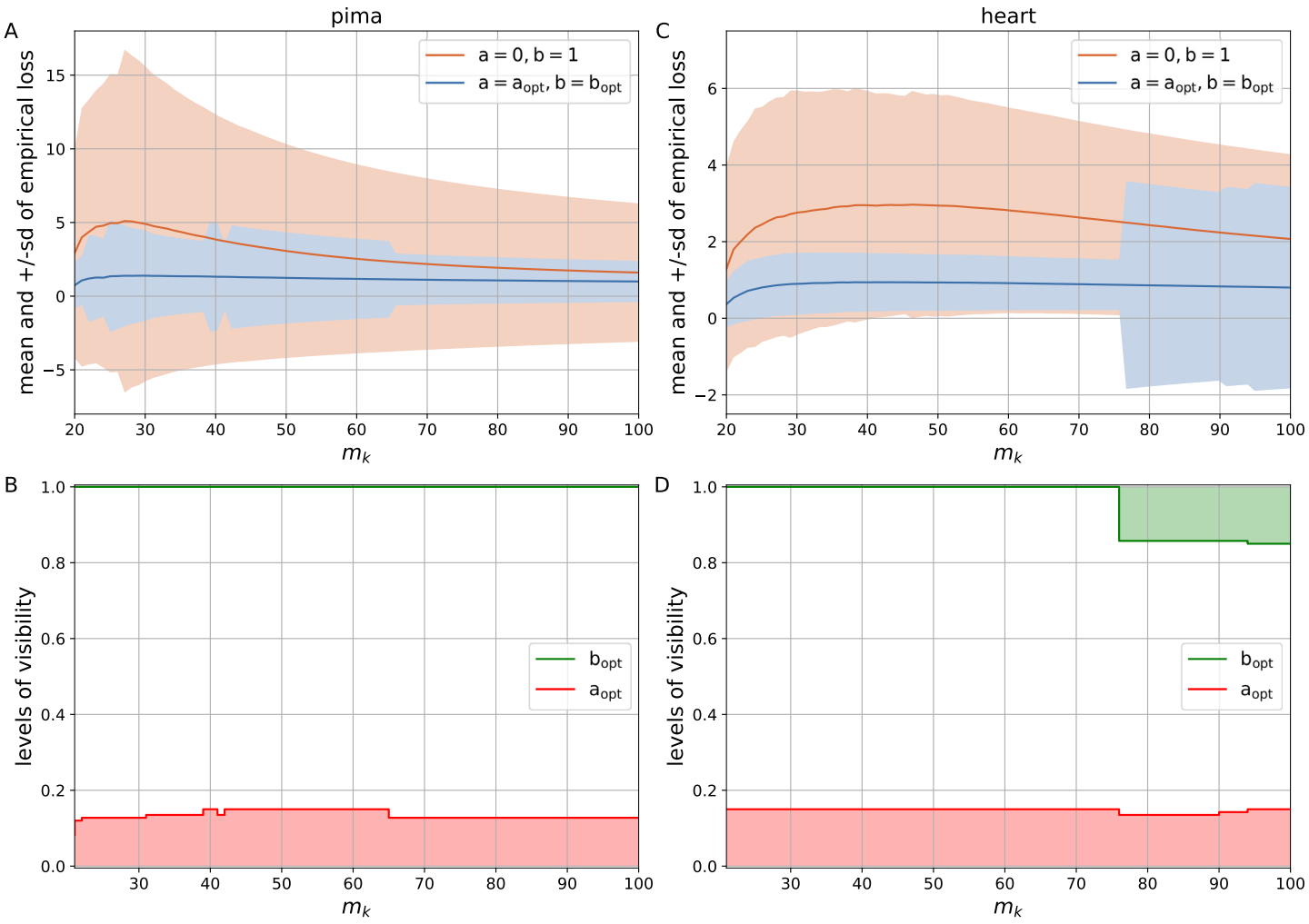

Fig 5. The estimates for the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to $m_k \in [20, 100]$. (A) The mean and +/-sd of empirical loss for the *pima* dataset. (B) The levels of visibility a_{opt}, b_{opt} for the *pima* dataset. (C) The mean and +/-sd of empirical loss for the *heart* dataset. (D) The levels of visibility a_{opt}, b_{opt} for the *heart* dataset.

https://doi.org/10.1371/journal.pone.0219004.g005

introduced the hyperparameters a, b, and r into the model. We used the default values for the Adam optimizer: $\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The neural network was trained on the freely available MNIST handwritten digits dataset [56] to solve a 10-class classification problem. The dataset consists of 10 handwritten digits {0, .., 9} from 250 different people, 50% of high school students, and 50% of employees from the Census Bureau. Each feature vector contains $n = 784$ pixels unrolled from the original 28x28 pixels images. The dataset consisted of $m = 60000$ examples for training and $m = 10000$ examples for testing. The training set was divided into an actual training set of 50000 examples and 10000 validation examples for selecting the hyperparameters. The number of epochs and the batch size for training were $n_{epoch} = 1$ and $n_{batch} = 25$, respectively. The hyperparameters were optimized with a random search as a less computationally expensive alternative to a grid search in the hyperparameter space $a \times b \times r$, where $a \in [0, 0.15]$, $b \in [0.85, 1]$, and $r \in [0.5, 5.1]$. We tested the model on a subset with adversarial perturbations to confirm the importance of the hyperparameters for regulating the degree to which changes in gradients impact prediction.

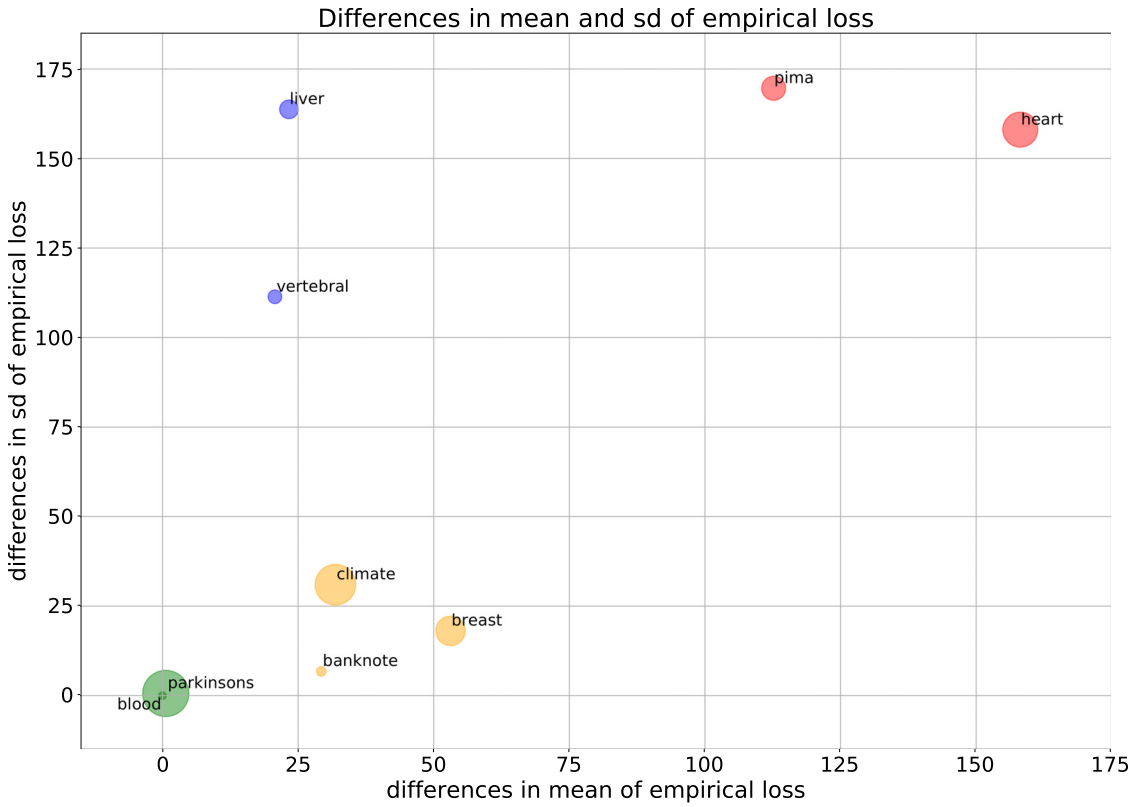

Fig 6. The differences in mean and sd of empirical loss between the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} for all the datasets summed over the interval $m_k \in [20, 100]$.

https://doi.org/10.1371/journal.pone.0219004.g006

The model of neural network was built in TensorFlow using Keras API. We also implemented a random search with a random pick 15% of the permutations for optimizing the hyperparameters with Talos and FGSM [11] for generating gradient-based attacks with Foolbox [57]. FGSM allowed us to add the signs of gradients to the images, and, by that, increase the magnitude until the images were misclassified.

Results. Fig 7 depicts the probabilities of predicting correct images on a testing dataset in presence of adversarial noise for the advanced sigmoid activation function and the simple activation function. We can see that the advanced function with the optimal hyperparameters did not bring distinct advantages (see Fig 7A) but still allowed us to guarantee higher accuracy that is 95.6% in comparison with the simple function that gives 92.8% on a testing dataset without adversarial examples.

By definition, the hyperparameter a allows us to regulate the magnitude of gradients. Consequently, we set a threshold value $a_{threshold} = 0.13$ that was used on training the neural network model. The neural network model with the parameter $a_{threshold}$ instead of a_{opt} still showed better prediction on a testing dataset that does not contain adversarial noise—94.6% against 92.8%. We investigated the influence of this parameter on the probability of predicting correct labels. The values b and r were fixed for clarity. Fig 7B) reflects a more marked increase in the probabilities of predicting correct labels for the advanced activation function in comparison with the simple activation function.

Consequently, the proposed optimization technique may contribute to increasing robustness to adversarial noise by penalizing the degree to which subtle changes in gradients impact prediction.

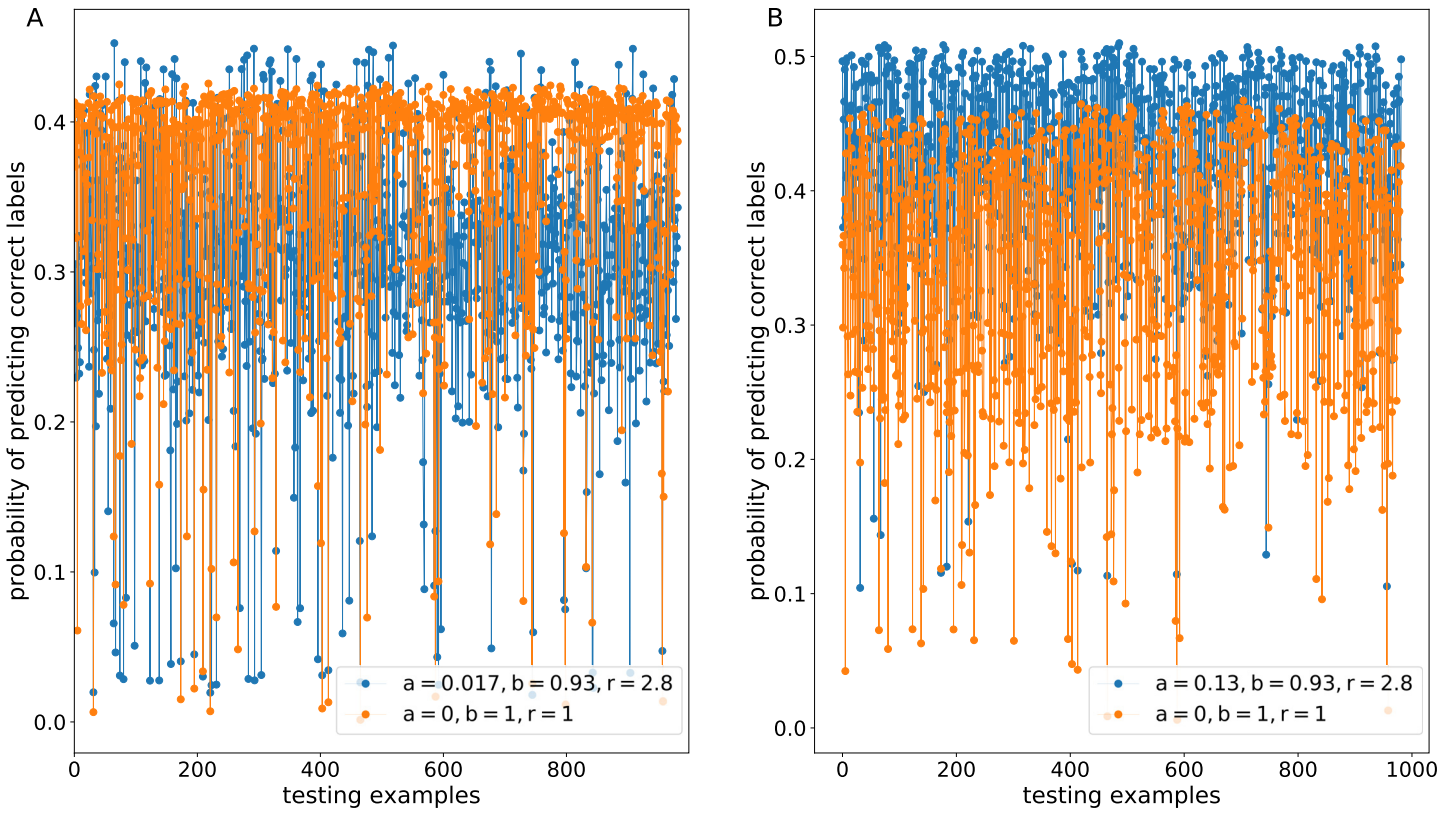

Fig 7. The probabilities of predicting correct labels in the presence of adversarial perturbations for the advanced sigmoid activation function with the hyperparameters a_{opt}, b_{opt}, r_{opt} and the simple sigmoid activation function with the hyperparameters a = 0, b = 1, r = 1. (A) The hyperparameters are equal to a_{opt} = 0.017, b_{opt} = 0.93, r_{opt} = 2.8. (B) The hyperparameters are equal to $a_{threshold}$ = 0.13, b_{opt} = 0.93, r_{opt} = 2.8.

https://doi.org/10.1371/journal.pone.0219004.g007

Discussion

BioGD shares some similarity with the gradient regularization method presented by Ross and Doshi-Velez [12] which allows regulating how much infinitesimal noise in gradients affect predictions. However, Ross and Doshi-Velez proposed a second-order method which implies a cost of taking second derivatives. Yu et al. [13], in turn, proposed a novel robust training method by regulating the first-order gradients of neural networks. But, the adversarial examples and neural networks were considered as quasi-linear models.

The BioGD is the first-order gradient method which aims to solve a predefined optimization problem, but also involves an open-ended adaptation process with regard to two hyperparameters. Considering the fact that these hyperparameters directly deal with vulnerable gradients, are adaptive to the newly updated data, and allow us to improve robustness to adversarial noise by regularizing the degree to which hardly visible changes in gradients impact prediction, the BioGD algorithm seems to be an efficient solution to the posed problem.

There are two important limitations to this research. First, the BioGD is based on the Verhulst logistic growth equation. An alternative model of logistic growth is the Gompertz equation that shows more flexible behavior near the lower and upper asymptotes in comparison with the Verhulst equation. Consequently, it seems promising to explore the effect of the asymptotes of the Gompertz equation on the magnitude of gradients below the lower and upper levels of visibility.

The other limitation is related to the open-ended adaptation process of the pair of hyperparameters. The implemented hyperparameter optimization controls the behavior of a learning algorithm using the empirical loss on a validation subset. Optimizing the hyperparameters on the same dataset tunes only the values of this pair. This adaptation process may be interpreted in terms of the single species population growth. As a direction for future research, we suggest adapting BioGD to a meta-learning framework that infers a learning algorithm from a set of datasets in order to improve robustness on unseen datasets. In contrast to hyperparameter optimization, meta-learning would imply optimizing a set of the pairs of hyperparameters and may be considered in the context of multiple species population growth.

Supporting information

S1 File. The estimates of mean and +/-sd of empirical loss subject to $m_k \in [20, 100]$ and $p = 0.0$.
(CSV)

S2 File. The estimates of mean and +/-sd of empirical loss subject to $m_k \in [20, 100]$ and $p = 0.3$.
(CSV)

S1 Fig. The estimates for the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to $m_k \in [20, 100]$ and $p = \{0.05, 0.1\}$. (A) The mean and +/-sd of empirical loss subject to $p = 0.05$. (B) The levels of visibility a_{opt}, b_{opt} subject to $p = 0.05$. (C) The mean and +/-sd of empirical loss subject to $p = 0.1$. (D) The levels of visibility a_{opt}, b_{opt} subject to $p = 0.1$.
(EPS)

S2 Fig. The estimates for the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to $m_k \in [20, 100]$ and $p = \{0.2, 0.25\}$. (A) The mean and +/-sd of empirical loss subject to $p = 0.2$. (B) The levels of visibility a_{opt}, b_{opt} subject to $p = 0.2$. (C) The mean and +/-sd of empirical loss subject to $p = 0.25$. (D) The levels of visibility a_{opt}, b_{opt} subject to $p = 0.25$.
(EPS)

S3 Fig. The estimates for the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to $m_k \in [20, 100]$. (A) The mean and +/-sd of empirical loss for the *banknote* dataset. (B) The levels of visibility a_{opt}, b_{opt} for the *banknote* dataset. (C) The mean and +/-sd of empirical loss for the *breast-win* dataset. (D) The levels of visibility a_{opt}, b_{opt} for the *breast-win* dataset. (E) The mean and +/-sd of empirical loss for the *climate* dataset. (F) The levels of visibility a_{opt}, b_{opt} for the *climate* dataset.
(EPS)

S4 Fig. The estimates for the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to $m_k \in [20, 100]$. (A) The mean and +/-sd of empirical loss for the *blood* dataset. (B) The levels of visibility a_{opt}, b_{opt} for the *blood* dataset. (C) The mean and +/-sd of empirical loss for the *parkinsons* dataset. (D) The levels of visibility a_{opt}, b_{opt} for the *parkinsons* dataset.
(EPS)

S5 Fig. The estimates for the simple logistic loss with a = 0, b = 1 and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to $m_k \in [20, 100]$. (A) The mean and +/-sd of empirical loss for the *chronic-kidney* dataset. (B) The levels of visibility a_{opt}, b_{opt}

for the *chronic-kidney* dataset. (C) The mean and +/-sd of empirical loss for the *cervical* dataset. (D) The levels of visibility a_{opt}, b_{opt} for the *cervical* dataset.
(EPS)

S6 Fig. The estimates for all the datasets summed over the interval $m_k \in [20, 100]$. (A) The mean of empirical loss against the sd of empirical loss for the simple logistic loss with a = 0, b = 1. (B) The mean of empirical loss against the sd of empirical loss for the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt}.
(EPS)

Acknowledgments

The authors would like to thank the editor and anonymous reviewers for valuable comments and suggestions.

Author Contributions

Conceptualization: Ilona Kulikovskikh.

Formal analysis: Ilona Kulikovskikh.

Funding acquisition: Ilona Kulikovskikh, Sergej Prokhorov, Tomislav Šmuc.

Investigation: Ilona Kulikovskikh.

Methodology: Ilona Kulikovskikh.

Project administration: Sergej Prokhorov, Tomislav Šmuc.

Resources: Sergej Prokhorov, Tomislav Šmuc.

Software: Ilona Kulikovskikh.

Supervision: Sergej Prokhorov, Tomislav Šmuc.

Validation: Ilona Kulikovskikh, Tomislav Lipić, Tarzan Legović, Tomislav Šmuc.

Visualization: Ilona Kulikovskikh, Tomislav Lipić.

Writing – original draft: Ilona Kulikovskikh.

Writing – review & editing: Tomislav Lipić, Tarzan Legović, Tomislav Šmuc.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521: 436–444. https://doi.org/10.1038/nature14539 PMID: 26017442
2. Cherry KM, Qian L. Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks. Nature. 2018; 559: 370–376. https://doi.org/10.1038/s41586-018-0289-6 PMID: 29973727
3. Jang BS, Jeon SH, Han Kim IIH, Kim IA. Prediction of pseudoprogression versus progression using machine learning algorithm in glioblastoma. Scientific Reports. 2018; 8: 12516. https://doi.org/10.1038/s41598-018-31007-2 PMID: 30131513
4. Maxmen A. Deep learning sharpens views of cells and genes. Nature. 2018; 553: 9–10. https://doi.org/10.1038/d41586-018-00004-w PMID: 29300023
5. Nirschl JJ, Janowczyk A, Peyster EG, Frank R, Margulies KB, Feldman MD, et al. A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue. PLoS ONE. 2018; 13(4): e0192726. https://doi.org/10.1371/journal.pone.0192726 PMID: 29614076
6. Webb S. Deep learning for biology. Nature. 2018; 554: 555–557. https://doi.org/10.1038/d41586-018-02174-z PMID: 29469107

7. Shariff A, Bonnefon JF, Rahwan I. Psychological roadblocks to the adoption of self-driving vehicles. Nature Human Behaviour. 2017; 1: 694–696. https://doi.org/10.1038/s41562-017-0202-6 PMID: 31024097
8. Floreano D, Robert J. Wood Science, technology and the future of small autonomous drones. Nature. 2015; 521: 460–466. https://doi.org/10.1038/nature14542 PMID: 26017445
9. Dawes J. The case for and against autonomous weapon systems. Nature Human Behaviour. 2017; 1: 613–614. https://doi.org/10.1038/s41562-017-0182-6 PMID: 31024133
10. Taddeo M, Floridi L. Regulate artificial intelligence to avert cyber arms race. Nature. 2018; 556: 296–298. https://doi.org/10.1038/d41586-018-04602-6 PMID: 29662138
11. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. ICLR. 2015. Available from: https://arxiv.org/abs/1412.6572
12. Ross AS, Finale Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. AAAI. 2018: 1660-1669.
13. Yu F, Xu Z, Wang Y, Liu C, Chen X. Towards robust training of neural networks by regularizing adversarial gradients. arXiv 1805.09370 [Preprint]. 2018 [cited 2018 Dec 30]. Available from: https://arxiv.org/abs/1805.09370
14. Varga D, Csiszárik A, Zombori Z. Gradient regularization improves accuracy of discriminative models. arXiv 1712.09936 [Preprint]. 2018 [cited 2018 Dec 30]. Available from: https://arxiv.org/abs/1712.09936
15. Smilkov D, Thorat N, Kim B, Vivegas F, Wattenberg M. SmoothGrad: removing noise by adding noise. ICML. 2018: 274–283.
16. Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. arXiv 1802.00420 [Preprint]. 2018 [cited 2018 Dec 30]. Available from: https://arxiv.org/abs/1802.00420
17. Biggioa B, Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition. 2018; 84: 317–331. https://doi.org/10.1016/j.patcog.2018.07.023
18. Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2574–2582.
19. Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. ICLR Workshop. 2018. Available from: https://arxiv.org/abs/1607.02533
20. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy. 2017: 39–57.
21. Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbation. IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1765–1773.
22. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. IEEE European Symposium on Security and Privacy. 2016: 372–387.
23. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv 1706.06083 [Preprint]. 2017 [cited 2018 Dec 30]. Available from: https://arxiv.org/abs/1706.06083.
24. Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks. 25th Annual Network and Distributed System Security Symposium. 2018. Available from: https://arxiv.org/abs/1704.01155.
25. Miyato T, Maeda S, Koyama M, Nakae K, Ishii S. Distributional smoothing with virtual adversarial training. ICLR. 2016. Available from: https://arxiv.org/abs/1507.00677.
26. Guo C, Rana M, Cisse M, van der Maaten L. Countering adversarial images using input transformations. ICLR. 2018. Available from: https://arxiv.org/abs/1711.00117.
27. Zantedeschi V, Nicolae MI, Rawat A. Efficient defenses against adversarial attacks. 10th ACM Workshop on Artificial Intelligence and Security. 2017. Available from: https://arxiv.org/abs/1707.06728.
28. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. ICLR. 2014. Available from: https://arxiv.org/abs/1312.6199.
29. Chen B, Carvalho W, Baracaldo N, Ludwig H, Edwards B, Lee T, et al. Detecting backdoor attacks on deep neural networks by activation clustering. Workshop on Artificial Intelligence Safety co-located with the 33rd AAAI Conference on Artificial Intelligence. 2019. Available from: https://arxiv.org/abs/1811.03728.
30. Li Y, Bradshaw J, Sharma Y. Are generative classifiers more robust to adversarial attacks? arXiv 1802.06552 [Preprint]. 2018 [cited 2018 Dec 30]. Available from: https://arxiv.org/abs/1802.06552.
31. Ilyas A, Jalal A, Asteri E, Daskalakis C, Dimakis AG. The robust manifold defense: Adversarial training using generative models. arXiv 1712.09196 [Preprint]. 2018 [cited 2018 Dec 30]. Available from: https://arxiv.org/abs/1712.09196.

32. Drucker H, Le Cun Y. Improving generalization performance using double backpropagation. IEEE Transactions on Neural Networks. 1992; 3(6): 991–997. https://doi.org/10.1109/72.165600 PMID: 18276495

33. Floreano D, Mattiussi C, Arkin RC. Bio-inspired artificial intelligence: Theories, methods, and technologies (Intelligent Robotics and Autonomous Agents series). Cambridge, Massachusetts: The MIT Press; 2008.

34. Ghodrati M, Khaligh-Razavi S-M, Ebrahimpour R, Rajaei K, Pooyan M. How can selection of biologically inspired features improve the performance of a robust object recognition model? PLoS ONE. 2012; 7 (2): e32357. https://doi.org/10.1371/journal.pone.0032357 PMID: 22384229

35. Mocanu DC, Mocanu E, Stone P, Nguyen PH, Gibescu M, Liotta A. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. Nature Communications. 2018; 9: 2383. https://doi.org/10.1038/s41467-018-04316-3 PMID: 29921910

36. Barradas-Bautista D, Alvarado-Mentado M, Agostino M, Cocho G. Cancer growth and metastasis as a metaphor of Go gaming: An Ising model approach. PLoS ONE. 2018. 13(5): e0195654. https://doi.org/10.1371/journal.pone.0195654 PMID: 29718932

37. Brodbeck L, Hauser S, Iida F. Morphological evolution of physical robots through model-free phenotype development. PLoS ONE. 2015. 10(6): e0128444. https://doi.org/10.1371/journal.pone.0128444 PMID: 26091255

38. Crandall JW, Oudah M, Tennom, Ishowo-Oloko F, Abdallah S, Bonnefon JF, et al. Cooperating with machines. Nature Communications. 2018; 9: 233. https://doi.org/10.1038/s41467-017-02597-8 PMID: 29339817

39. Cully A, Clune J, Tarapore D, Mouret JB. Robots that can adapt like animals. Nature. 2015; 521: 503–507. https://doi.org/10.1038/nature14422 PMID: 26017452

40. Floreano D, Keller L. Evolution of adaptive behaviour in robots by means of Darwinian Selection. PLoS Biol. 2010; 8(1): e1000292. https://doi.org/10.1371/journal.pbio.1000292 PMID: 20126252

41. Palagi S, Fischer P. Bioinspired microrobots. Nature Reviews Materials. 2018; 3: 113–124. https://doi.org/10.1038/s41578-018-0016-9

42. Yao Y, Marchal K, Van de Peer Y. Improving the adaptability of simulated evolutionary swarm robots in dynamically changing environments. PLoS ONE. 2014; 9(3): e90695. https://doi.org/10.1371/journal.pone.0090695 PMID: 24599485

43. Verhulst PF. Notice sur la loi que la population poursuit dans son accroissement. Correspondance mathamatique et physique. 1838; 10: 113–121.

44. Legović T. Dynamic population models. In ecological model types (ed. Jorgensen SE). Elsevier; 2016: 39–63.

45. Soudry D, Hoffer E, Nacso MS, Gunasekar S, Srebro N. The implicit bias of gradient descent on separable data. JMLR. 2018; 19: 1–57.

46. Foerster HV, Mora PM, Amiot LW. Doomsday: Friday, 13 November, A.D. 2026. Science. 1960; 132 (3436): 1291–1295. https://doi.org/10.1126/science.132.3436.1291

47. Malthus TR. An essay on the principle of population, as it affects the future improvement of society. With Remarks on the speculations of Mr. Godwin, M. Condorcet and other writers. London: J. Johnson; 1798.

48. Allee WC. Animal aggregations. The Quarterly Review of Biology. 1927; 2(3): 367–398. https://doi.org/10.1086/394281

49. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. Journal of Machine Learning Research. 2012; 13: 281–305.

50. Martinez-Cantin R, Tee K, McCourt M. Practical Bayesian optimization in the presence of outliers. PMLR. 2018; 84: 1722–1731.

51. Okun H, Persad S, Xia A. 6.854-Final-Project [Internet]. Github; Available: https://github.com/haroku/6.854-Final-Project.

52. Zhu X, Wu X. Class noise vs. Attribute noise: A quantitative study. Artificial Intelligence Review. 2004. 22(3): 177–210. https://doi.org/10.1007/s10462-004-0751-8

53. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin. 1945; 1(6): 80–83. https://doi.org/10.2307/3001968

54. McDonald JH. Handbook of Biological Statistics (3rd ed.). London: J. JohnsonSparky House Publishing, Baltimore, Maryland; 2014.

55. Kingma D, Ba J. Adam: A method for stochastic optimization. ICLR. 2015. Available from: https://arxiv.org/abs/1412.6980.

56. LeCun Y,Cortes C. Mnist handwritten digit database. AT&T Labs [Preprint]. 2010 [cited 2019 Mar 24]. Available from: http://yann.lecun.com/exdb/mnist.

57. Rauber J, Brendel W, Bethge M. Foolbox: A Python toolbox to benchmark the robustness of machine learning models. arXiv 1707.04131 [Preprint]. 2018 [cited 2019 Mar 24]. Available from: https://arxiv.org/abs/1707.04131.