24 October 2019

Robust machine learning inspired by the models of population dynamics

Ilona Kulikovskikh, Tomislav Šmuc

Outline

Problem statement

We consider a dataset $\{x_i, y_i\}_{i=1}^m$ with $x_i \in R^n$, $y_i \in \{0, 1\}$ and minimize an empirical loss function

$$\mathscr{L}(\theta) = \sum_{i=1}^{m} \ell(\theta^T x_i),$$

with a weight vector $\theta \in R^n$. We are interested in linearly separable problems with a smooth monotone strictly decreasing and non-negative loss function.

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

$$\theta_{I+1} = \theta_I - \eta \nabla \mathscr{L}(\theta_I) = \theta_I - \eta \sum_{i=1}^{m} \ell'(\theta_I^T x_i) x_i. \tag{1}$$

In equation (1) it is assumed that all the labels are positive:
$\forall i : y_i = 1, \|x_i\| < 1$, the dataset is linearly separable: $\exists\ \theta^*$ such that
$\forall i : \theta^{*T} x_i > 0$ and $\forall t$: $\ell(t)$ is differentiable and monotonically decreasing to zero and its derivative is β-Lipshitz.

Outline

Population dynamics models

The population growth $P(t)$ in an infinite environment can be described as:

$$\frac{dP(t)}{dt} = rP(t), \tag{2}$$

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 presents the following exponential growth model:

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

where P_0 is the initial population.

An infinite growth $\lim_{t\to\infty} P(t) = \infty$ of population does not consider the scarcity of resources. The Verhulst equation introduces the logistic growth model:

$$\frac{dP(t)}{dt} = rP(t)\left(1 - P(t)/K\right), \tag{4}$$

where K is the carrying capacity that represents the maximum size of population which can be supported by the environment.

If $\lim_{t\to\infty} P(t) = K$, the population stops growing.

Population dynamics models

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

$$\frac{dP(t)}{dt} = r\big(P(t) - A\big)\left(1 - \left(\frac{P(t) - A}{K - A}\right)\right); \tag{5}$$

$$P(t) = A + \frac{K - A}{1 - \left(1 - \left(\frac{K-A}{P_0+A}\right)\right)\exp(-rt)}, \tag{6}$$

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.

Population dynamics models

Let us introduce the generalized logistic equation and its solution:

$$\frac{dP(t)}{dt} = -r\,(P(t) - A)\ln^{[q]}\left(\frac{P(t) - A}{K - A}\right), \tag{7}$$

where

$$\ln^{[q]}(x) = \int_1^x \frac{dt}{t^{1-q}} = \begin{cases} \frac{x^q - 1}{q}, & q \neq 0; \\ \ln(x), & q \to 0, \end{cases} \tag{8}$$

$$P(t) = A + \frac{K - A}{\left(1 - \left(1 - \left(\frac{K-A}{P_0 + A}\right)^q\right)\exp(-rt)\right)^{\frac{1}{q}}}, \tag{9}$$

where q is the generalization parameter, $0 \leq A < K$, K is the upper asymptote or the carrying capacity of population; A is the lower asymptote that indicates critical population thresholds below which a population crashes to extinction.

Population dynamics models

Given that:

1. $q = 1$ and $K \to \infty$: the generalized model reduces to the Malthus model that describes the exponential growth of a population;
2. $q = 1$: the generalized model reduces to the Verhulst model that describes the logistic growth of a population;
3. $q \to 0$: the generalized model reduces to the Gompertz model that also describes the logistic growth of a population, but it is more flexible in the way of approaching the upper and lower asymptotes.

Population dynamics models

A discrete-time population model (the logistic map) at the time k is given as:

$$\Delta P_k = rP_k(1 - P_k),$$

where $P_k \in [0, 1]$, $k \in N$ and $r \in (0, 4]$. Let $\Delta t = 1$, $\ell'_r(P) = rP(1 - P)$. Then $P_{k+1} = \ell'_r(P_k)$ and the composition of k functions $\ell_r(P)$ can be represented by:

$$\mathcal{L}_r^{k'}(P) = \begin{cases} \ell'_r(P), & k = 1; \\ \left(\ell'_r \circ \mathcal{L}_r^{[k-1]'}\right)(P), & k > 1. \end{cases} \tag{10}$$

The equation (10) describes the dynamics of a population.

1. If the growth rate $0 < r \leq 1$, the population dies out and goes extinct.
2. Increasing the rate of growth allows the population to settle at the stable value or fluctuate across booms and busts.
3. Finally, at a relatively high values of growth rate, the logistic equation produces chaos.

Outline

Robustness to small variations in data

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

$$\ell_r'(t; a, b) = r(\ell_r(t; a, b) - a)\left(1 - \frac{\ell_r(t; a, b) - a}{b - a}\right) \tag{11}$$

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

$$\ell_r(t; a, b) = a + \frac{b - a}{\left(1 - \left(1 - \left(\frac{b-a}{P_0 + a}\right)\right)\exp(-rt)\right)}, \tag{12}$$

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

If $a = 0$, $b = 1$, and $r = 1$, then $\ell_1(t; 0, 1) = \ell(t)$.

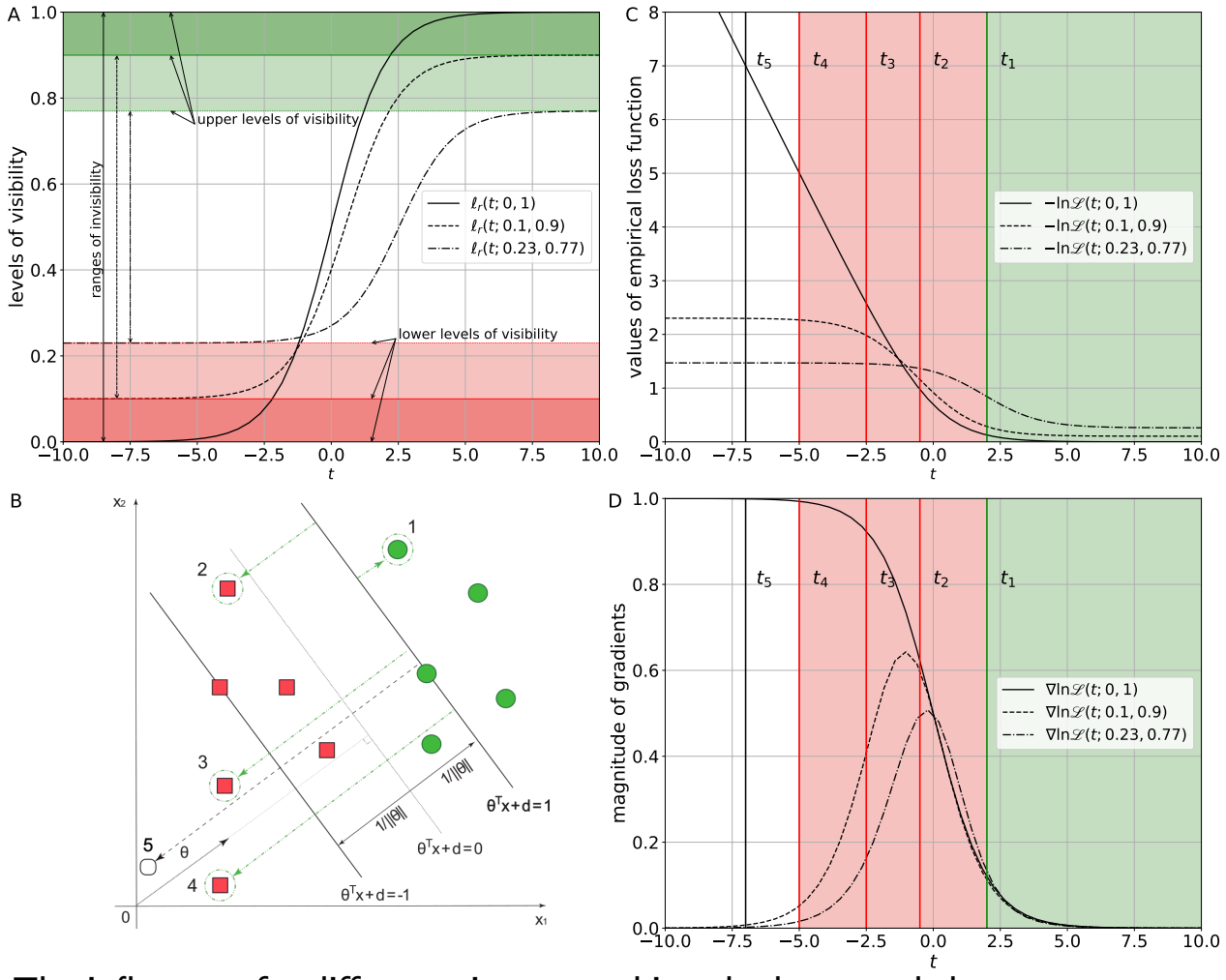


Figure 1: The influence of a difference in approaching the lower and the upper asymptote on the magnitude of gradients.
(A) The definitions of levels of visibility and ranges of invisibility based on the generalized logistic loss function $\ell_r(t; a, b)$ subject to the different pairs $(a, b) = \{(0, 1), (0.1, 0.9), (0.23, 0.77)\}$ 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 $-\ln\mathscr{L}(t; a, b)$ with the levels of visibility for reliable $t_{\mathbf{1}}$, noisy $t_{\mathbf{2}}$, $t_{\mathbf{3}}$, $t_{\mathbf{4}}$, and adversarial $t_{\mathbf{5}}$ 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 $\nabla \ln\mathscr{L}(t; a, b)$ subject to the different pairs $(a, b) = \{(0, 1), (0.1, 0.9), (0.23, 0.77)\}$ and $r = 1$.

Robustness to small variations in data

Applying (11) to the updates of gradient descent:

$$\theta_{l+1} = \theta_l - \eta \sum_{i=1}^{m} \ell_r'(\theta_l^T x_i; a, b) x_i = \theta_l - \eta_r \sum_{i=1}^{m} \ell'(\theta_l^T x_i) x_i, \tag{13}$$

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

The empirical generalized logistic loss function and its gradient $\forall y_i \in \{0, 1\}$ can be represented as follows:

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

$$\nabla \ln \mathscr{L}(\theta; a, b) = -\sum_{i=1}^{m} (y_i - \ell_r(\theta^T x_i; a, b)) \frac{\ell_r'(\theta^T x_i; a, b)}{(1 - \ell_r(\theta^T x_i; a, b))\ell_r(\theta^T x_i; a, b)} x_i.$$

Taking into account our assumption that $y_i = 1$, (13) can be given as:

$$\theta_{l+1} = \theta_l - \eta \sum_{i=1}^{m} \frac{\ell_r'(\theta_l^T x_i; a, b)}{\ell_r(\theta_l^T x_i; a, b)} x_i = \theta_l - \eta_r \sum_{i=1}^{m} \frac{\ell'(\theta_l^T x_i)}{\ell_r(\theta_l^T x_i; a, b)} x_i. \tag{14}$$

Robustness to small variations in data

Algorithm Bio-inspired gradient descent

procedure BioGD(x, y, η, n)
  Initialize $\boldsymbol{\theta}_0$;
  Initialize $a \in [a_{\min}, a_{\max}]$, $b \in [b_{\min}, b_{\max}]$, $r \in [r_{\min}, r_{\max}]$;
  Initialize a grid of n points in the space $a \times b \times r$;
  Split (x, y) into train $(\mathrm{x}, y)_T$ and cross-validation $(\mathrm{x}, y)_{CV}$ subsets;
  $l \leftarrow 0$;
  repeat
    $\boldsymbol{\theta}_{l+1} \leftarrow \boldsymbol{\theta}_l - (b-a) r \eta \nabla_{(\boldsymbol{\theta})} \mathscr{L}(\boldsymbol{\theta}_l, (\mathrm{x}, y)_T)$;
    $l \leftarrow l + 1$;
  until converge
  $(a, b, r) \leftarrow GridSearch(\mathscr{L}(\boldsymbol{\theta}_{l+1}, (\mathrm{x}, y)_{CV}), a, b, r)$;
  return $\boldsymbol{\theta}_{l+1}, (a, b, r)$

Robustness to small variations in data

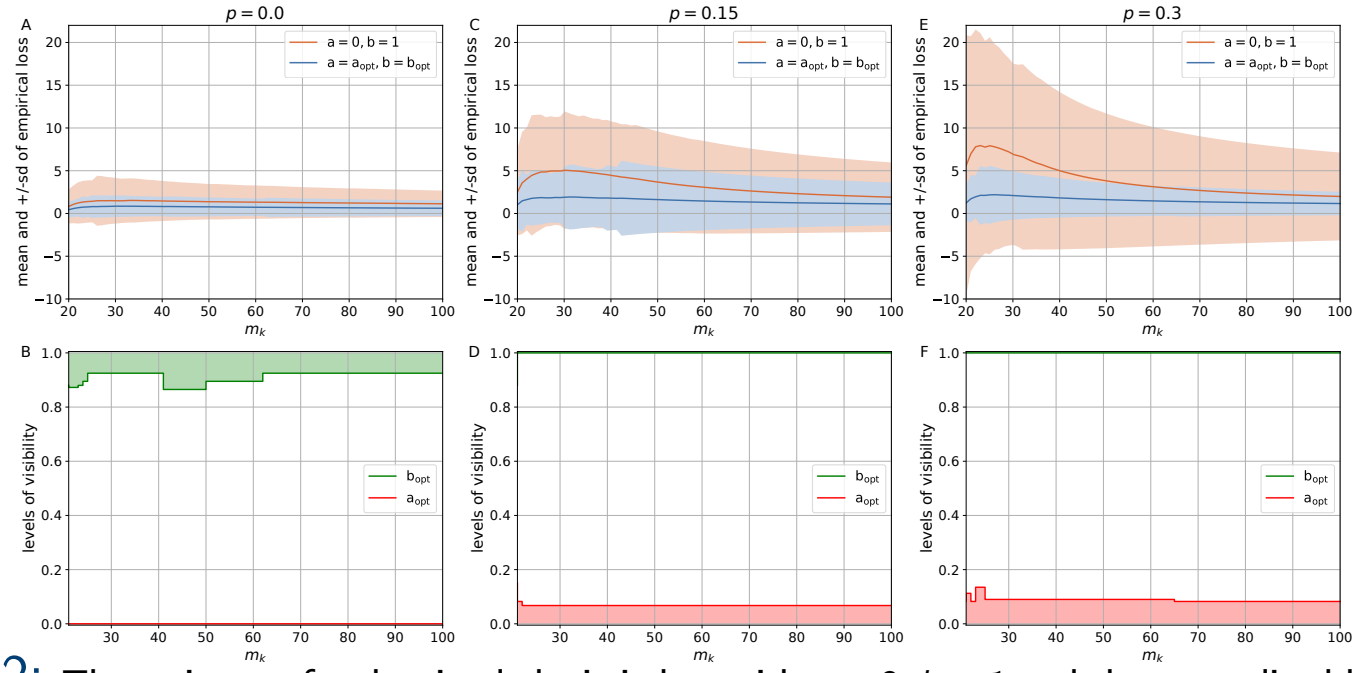


Figure 2: The estimates for the simple logistic loss with $a = 0$, $b = 1$ and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to $m_k \in [20, 100]$ and $p = \{0.0, 0.15, 0.3\}$. (A) The mean and +/-sd of empirical loss subject to $p = 0.0$. (B) The levels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

Robustness to small variations in data

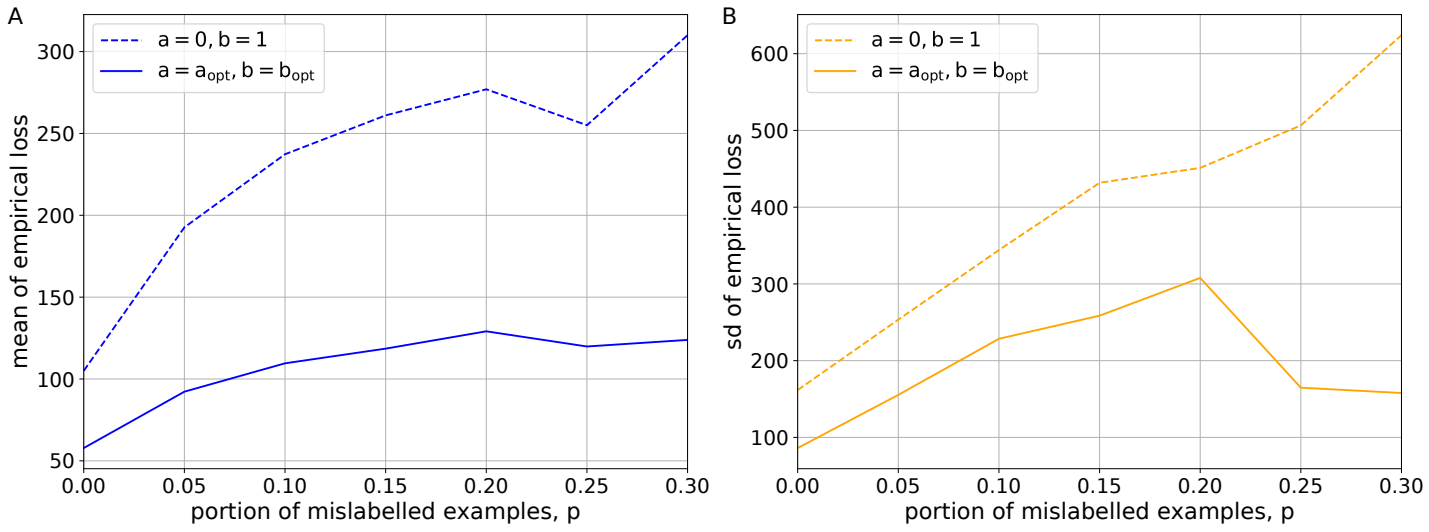


Figure 3: The estimates for the simple logistic loss with $a = 0, b = 1$ and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to the proportions of noisy labels $p = \{0.0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$.
(A) The mean of empirical loss summed over the interval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

Robustness to small variations in data

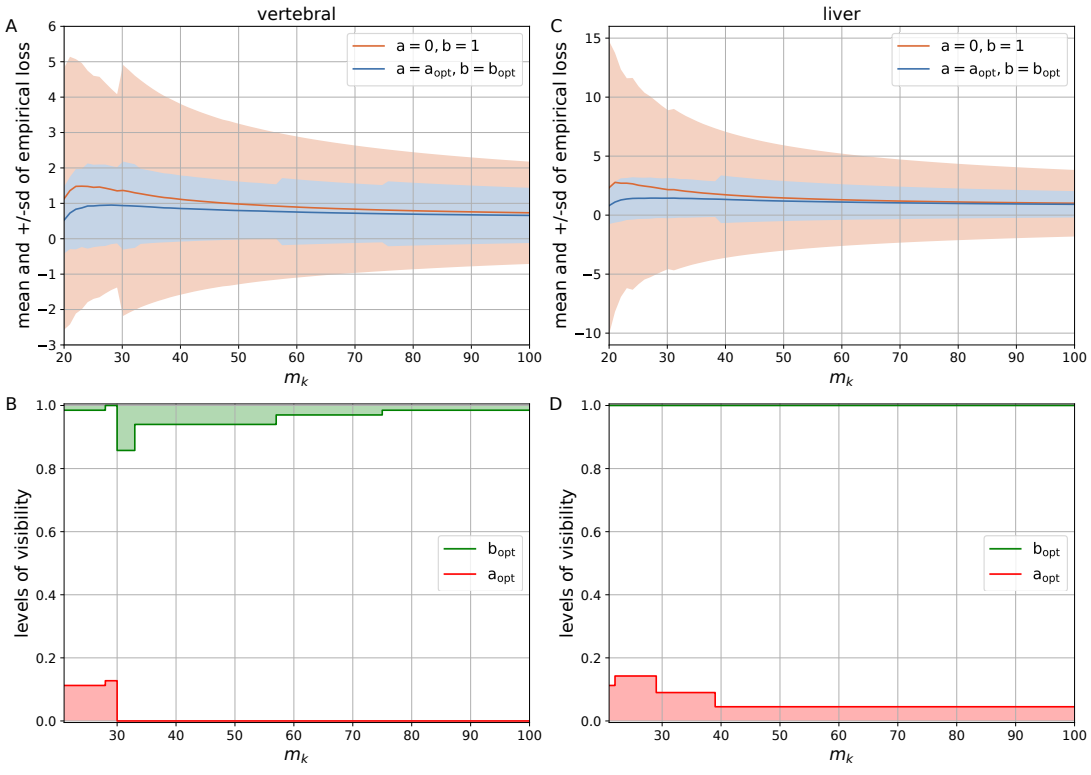


Figure 4: The estimates for the simple logistic loss with $a = 0, b = 1$ and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} subject to $m_k \in [20, 100]$.
(A) The mean and +/-sd of empirical loss for the *vertebral* dataset. (B) The levels of visibility a_{opt},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

Robustness to small variations in data

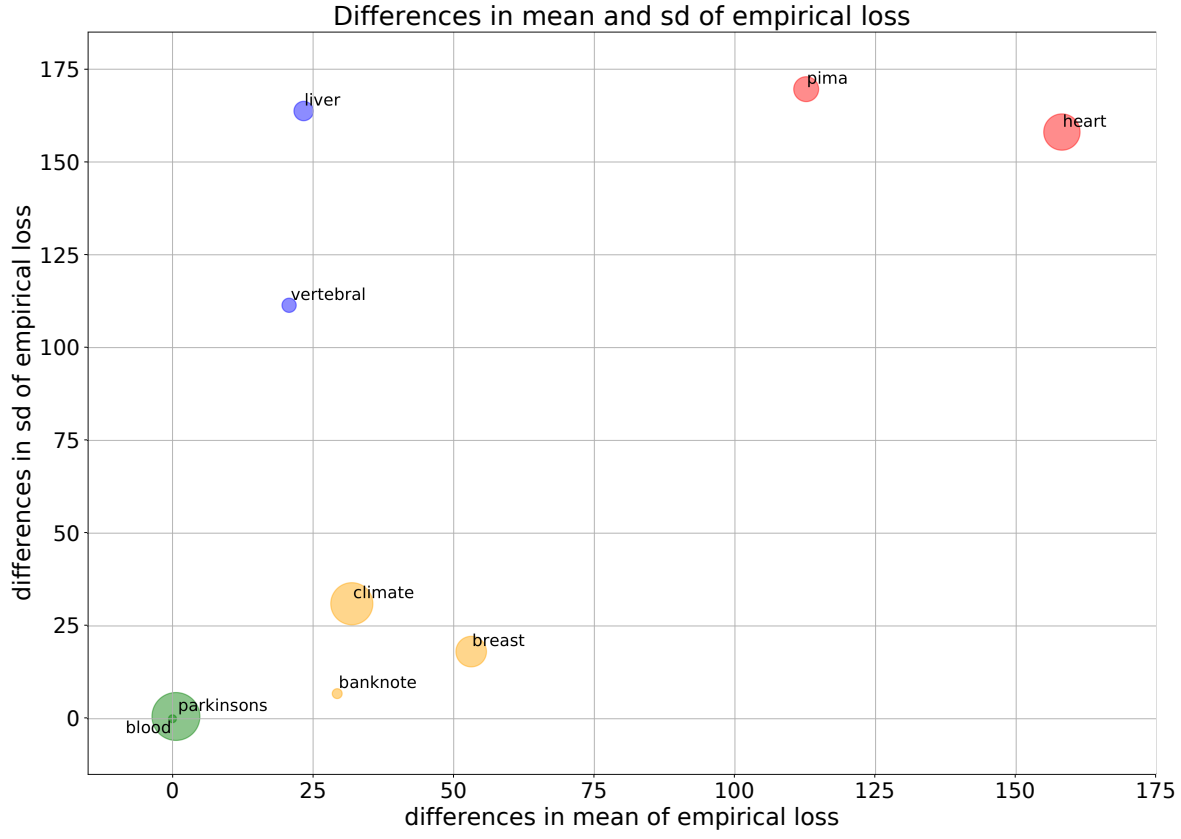


Figure 5: The differences in mean and sd of empirical loss between the simple logistic loss with $a = 0, b = 1$ and the generalized logistic loss with the optimal levels of visibility a_{opt}, b_{opt} for all the datasets summed over the interval $m_k \in [20, 100]$.

Outline

Meta-learning on almost separable data

The empirical logistic loss and its derivative with regard to the discrete population models 10 can be given as follows:

$$\ln \mathscr{L}(\theta; x_i, \mathcal{L}_r^{'k}(a, b, q)) = -\sum_{l=1}^{i} \ln \ell(\theta^T x_l; a, b, q),$$

$$-\nabla \ln \mathscr{L}(\theta; x_i, \mathcal{L}_r^{'k}(a, b, q)) = -\sum_{l=1}^{i} \left(\left(1 - \ell(\theta^T x_l; a, b, q)\right) \frac{\mathcal{L}_r^{'k}(\theta^T x_l; a, b, q)}{\ell'_r(\theta^T x_l)} x_l \right),$$

where $\mathcal{L}_r^{'k}(\theta^T x_l; a, b, q)$ defines the generalized logistic map.
A SGD-based meta-learner with "growing" descent based on the generalized logistic map can be defined as

$$\theta_{t+1} = \theta_t - \eta_{r_j^k} \nabla \ln \mathscr{L}(\theta_t; x_i, \mathcal{L}_r^{'k}(a_j, b_j, q)),$$

$$(a, b, r^k)_{j+1} \leftarrow (a, b, r^k)_j - \eta \nabla \mathscr{L}(\theta_{t+1}; x_i, \mathcal{L}_r^{'k}(a_j, b_j, q)),$$

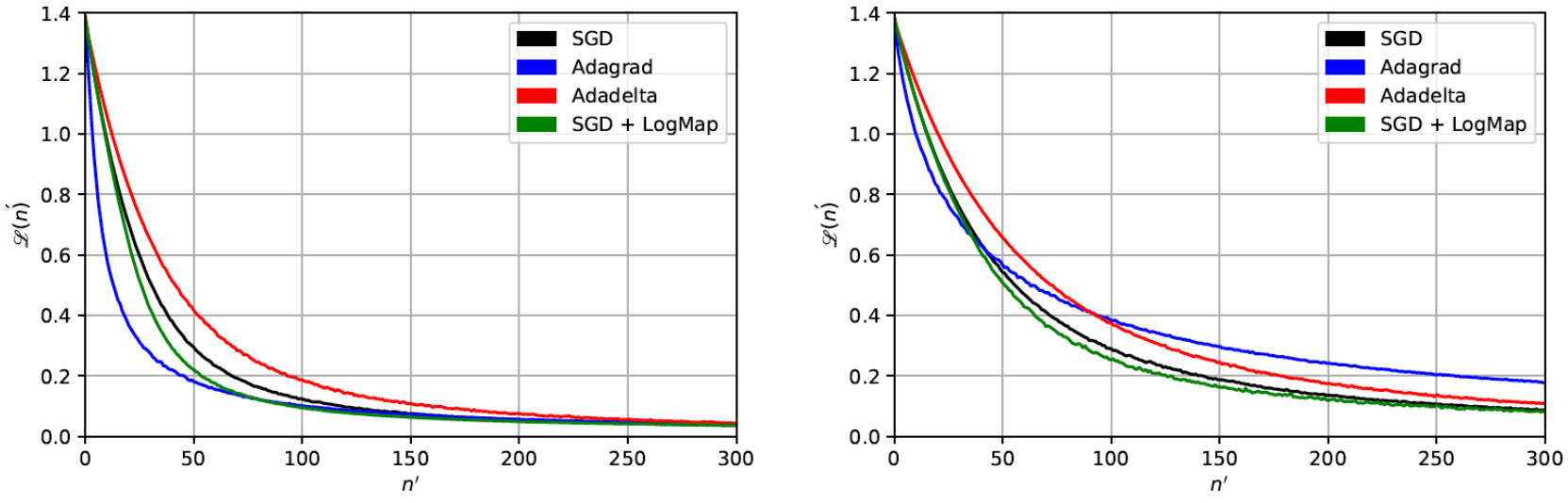
where the adaptive step size $\eta_{r_j^k} = (b - a)c(a_j, b_j)r_j^k \eta$.

Meta-learning on almost separable data

Algorithm 1 SGD-based meta-learner with "growing" descent

procedure GROWINGSGD(x, η, r)
  Initialize $\boldsymbol{\theta}_0$;
  $a_0 \leftarrow 0$, $b_0 \leftarrow 1$, $r_0^k \leftarrow 1$;
  Split x_i into train x_i^T and cross-validation x_i^{CV} subsets;
  $j \leftarrow 0$;
  repeat
    $t \leftarrow 0$;
    repeat
      $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta_{r_j^k} \nabla_{(\boldsymbol{\theta})} \mathscr{L}(\boldsymbol{\theta}_t; \mathrm{x}_i^{\mathrm{T}}, \mathcal{L}_r'^k(a_j, b_j, q))$;
      $t \leftarrow t + 1$;
    until converge
    $(a, b, r^k)_{j+1} \leftarrow (a, b, r^k)_j - \eta \nabla_{(a,b,r^k)} \mathscr{L}(\boldsymbol{\theta}_{t+1}; \mathrm{x}_i^{\mathrm{CV}}, \mathcal{L}_r'^k(a_j, b_j, q))$
    $j \leftarrow j + 1$;
  until converge
  return $\boldsymbol{\theta}_{t+1}, (a, b, r^k)_{j+1}$

Meta-learning on almost separable data



a) $n_{\text{neuron}} = 4,\ r = 2.8$ b) $n_{\text{neuron}} = 1,\ r = 3.8$

Figure 6: The results of computational experiments on a linear separable dataset modelled with $\mathcal{N}(0, 1)$ subject to $m = 5000$, $n = 10$, $n_{epochs} = 100$, the mini-batch size $m_{batch} = 100$ and the growth parameters $a = 0$, $b = 1$, $q = 1$

Meta-learning on almost separable data

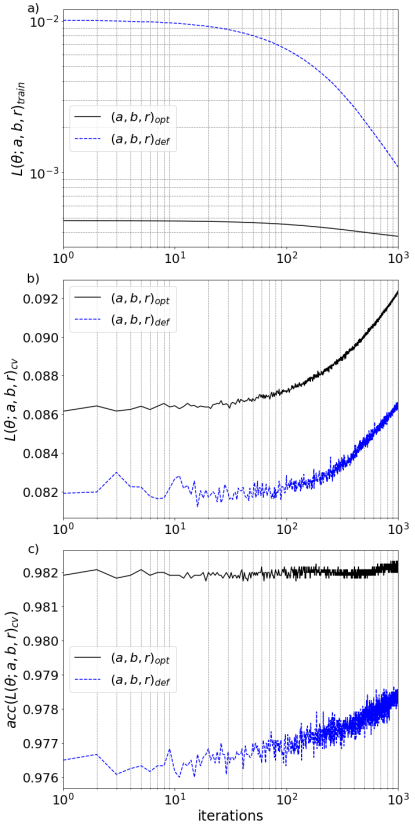


Figure 7: The model performance on the MNIST dataset for the simple and generalized logistic loss: a) training loss; b) validation loss; c) validation accuracy

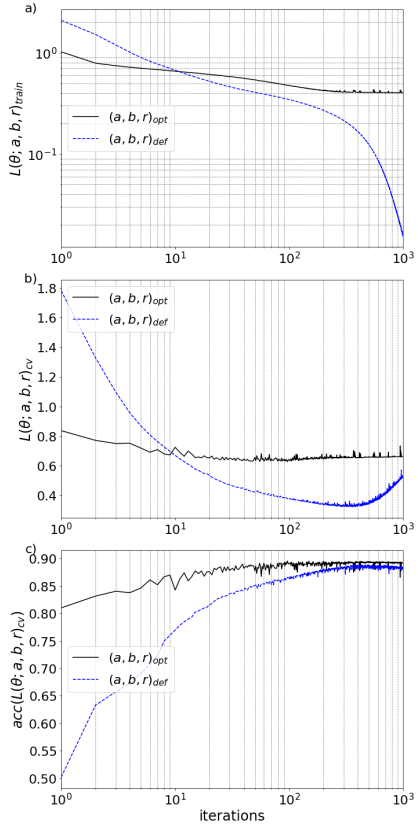


Figure 8: The model performance on the Fashion MNIST dataset for the simple and generalized logistic loss: a) training loss; b) validation loss; c) validation accuracy

Outline

Publications

Journal papers

1. Kulikovskikh I, Prokhorov S, Lipić T, Legović T, Šmuc T. BioGD: Bio-inspired robust gradient descent. 2019. PLoS ONE 14(7): e0219004.
2. Kulikovskikh I, Prokhorov S, Legović T, Šmuc T. An SGD-based meta-learner with "growing" descent. 2019. Journal of Physics: Conference Series. (accepted)
3. Kulikovskikh I. Reducing computational costs in deep learning on almost linearly separable training data. 2019. Computer Optics. (submitted)

Ongoing research

1. Accelerating the convergence of gradient descent on separable data with population dynamics models. 2019. Preprint.
2. Machines in a classroom: Towards human-like active learning. 2019. Preprint.