Снижение вычислительных затрат в глубоком обучении при почти идеальной линейной разделимости обучающей выборки

И.М. Куликовских [1,2,3]

[1] *Самарский национальный исследовательский университет имени академика С.П. Королёва, 443086, Россия, г. Самара, Московское шоссе 34,*

[2] *Факультет электротехники и вычислительной техники, Загребский университет, 10000, Хорватия, г. Загреб, Унска 3,*

[3] *Институт Руджер Бошкович, 10000, Хорватия, г. Загреб, Биеничка 54*

Аннотация

Последние исследования в области глубокого обучения показали, что метод градиентного спуска при условии почти идеальной разделимости обучающей выборки сходится к оптимальному решению, обеспечивающему максимальный зазор между классами. Даже без введения явной регуляризации положение разделяющей гиперплоскости продолжает изменяться, несмотря на то, что ошибка классификации на обучении стремится к нулю. Данное свойство так называемой «неявной» регуляризации позволяет использовать градиентный метод с более агрессивным шагом обучения, что гарантирует более низкие вычислительные затраты. Однако, хотя метод градиентного спуска обеспечивает хорошую обобщающую способность при стремлении к оптимальному решению, скорость сходимости к данному решению в условиях почти идеальной линейной разделимости значительно ниже, чем скорость сходимости, определяемая самой функцией потерь с заданным шагом обучения. В данной работе предлагается расширенная логарифмическая функция потерь, оптимизация параметров которой повышает скорость сходимости, обеспечивая границу погрешности, эквивалентную границе метода градиентного спуска. Результаты вычислительных экспериментов при классификации изображений на эталонных наборах MNIST и Fashion MNIST подтвердили эффективность предложенного подхода к снижению вычислительных затрат в условиях почти идеальной линейной разделимости обучающей выборки и обозначили направления дальнейших исследований.

Ключевые слова: неявная регуляризация, градиентный метод, скорость сходимости, линейная разделимость, классификация изображений.

Цитирование: **Куликовских, И.М.** Снижение вычислительных затрат в глубоком обучении при почти идеальной линейной разделимости обучающей выборки / И.М. Куликовских // Компьютерная оптика. – 2020. – Т. 44, № 2. – С. 282-289. – DOI: 10.18287/2412-6179-CO-645.

Citation: Kulikovskikh IM. Reducing computational costs in deep learning on almost linearly separable training data. Computer Optics 2020; 44(2): 282-289. DOI: 10.18287/2412-6179-CO-645.

Введение

В ряде последних исследований обнаружена важная особенность моделей глубокого обучения [1, 2], сводящая к минимуму ошибку обучения на почти идеально линейно разделимых выборках [3 – 9]. Без явной регуляризации модели с большим числом параметров часто демонстрируют хорошую способность к обобщению, поскольку итерации градиентного метода продолжают смещать разделяющую гиперплоскость к оптимальному положению, даже если ошибка классификации на обучении равна нулю [4]. Данное явление получило название «неявной» регуляризации [3, 4, 8, 9]. Свойство «неявной» регуляризации позволяет градиентному спуску проходить траекторию оптимизации более агрессивно, без перерегулирования, что, в свою очередь, приводит к значительной экономии вычислительных затрат.

Несмотря на очевидные преимущества наличия неявной регуляризации, скорость сходимости, определяемая самой функцией потерь с заданным шагом обучения, является линейной $O(1/t)$, тогда как скорость сходимости к оптимальному решению в условиях почти линейной разделимости классов лишь логарифмическая $O(1/\ln t)$ [7].

Наиболее часто используемый подход к повышению скорости сходимости заключается в применении методов оптимизации с переменным шагом, таких как Adam [10], Adagrad [11], Adadelta [12] и т. д. [13, 14]. Использование адаптивных шагов обучения снижает смещение, но приводит к ухудшению обобщающей способности [6, 13, 15]. Кроме того, направление оптимизации адаптивных методов менее предсказуемо в сравнении с неадаптивными методами [8].

В работе [7] исследовано влияние различных типов функций потерь на скорость сходимости. Согласс-

но результатам проведённого исследования, функции потерь с экспоненциальными хвостами достигают оптимальной скорости сходимости, равной $O(\ln t/\sqrt{t})$. В данной работе предлагается модификация логарифмической функции потерь, которая сводится к экспоненциальной и логистической функции потерь при заданных значениях гиперпараметров. Оптимизация данных параметров приводит к скорости сходимости, близкой к $O(\ln t/\sqrt{t})$ и $O(1/t)$, гарантируя границу погрешности метода градиентного спуска.

Данная статья изложена следующим образом. Параграф 1 посвящён математической постановке задачи. Параграф 2 описывает предлагаемый в работе подход к снижению вычислительных затрат в условиях почти линейной разделимости обучающей выборки. В параграфе 3 приведены результаты вычислительных экспериментов при классификации изображений. В заключении перечислены основные результаты, рекомендации по практическому использованию и дальнейшие направления исследований.

1. Математическая постановка задачи

Дана совокупность наблюдений

$$\{\mathrm{x}_i, y_i\}_{i=1}^m,$$

где $x_i \in \mathbb{R}^n$ и $y_i \in \{0,1\}$. Поставим задачу минимизации эмпирической функции потерь

$$L(\theta)=\sum_{i=1}^{m} l\left(y_i \theta^{\mathrm{T}} \mathrm{x}_i\right), \tag{1}$$

где $\theta \in \mathbb{R}^n$ задает вектор параметров модели. По аналогии с постановкой задачи в [4], для простоты представления сделаем предположение, что $\forall i \in \{1,...,m\}: y_i = 1$. Рассмотрим случай, когда выборка наблюдений *почти* идеально разделима, т.е. $\exists \theta^*$ такое, что $\forall i: \theta^{*\mathrm{T}} x_i > 0$, где лишь объекты-выбросы классифицируются неверно [16], а функция потерь l является гладкой строго убывающей неотрицательной функцией:

$$\forall t \in \mathbb{R}: l(t) > 0, l'(t) < 0, \lim_{t\to\infty} l(t) = \lim_{t\to\infty} l'(t) = 0, \tag{2}$$

имеющей непрерывный по Липшицу градиент с константой $\beta > 0$:

$$l(t') \le l(t) + \left\langle \nabla l(t), t'-t \right\rangle + \frac{\beta}{2}\left\|t'-t\right\|^2, \tag{3}$$

где $\lim\limits_{t\to-\infty} l'(t) \neq 0$.

Согласно Определению 2 в работе [7], отрицательная производная функция потерь $-l'(t)$ имеет экспоненциальный хвост, если существуют положительные константы c, a, μ_+, μ_-, t_+, t_-, такие, что:

$$\forall t > t_+ : l'(t) \ge c\left(1+\exp\left(-\mu_+ t\right)\right)\exp\left(-at\right),$$
$$\forall t > t_- : l'(t) \le c\left(1+\exp\left(-\mu_- t\right)\right)\exp\left(-at\right).$$

Определения (2) и (3) при различных значениях констант включают множество функций потерь, включая экспоненциальную и логарифмическую функции.

Решение задачи $\min\limits_{\theta\in\mathbb{R}^n} L(\theta)$ может быть найдено на j-й итерации метода градиентного спуска с шагом η:

$$\theta_{j+1} = \theta_j - \eta \nabla L\left(\theta_j\right) = \theta_j - \eta \sum_{i=1}^{m} l'\left(\theta_j^{\mathrm{T}} x_i\right) x_i. \tag{4}$$

В работах [4] было показано, что в условиях идеальной разделимости выборки наблюдений справедливо равенство:

$$\theta_t = \hat{\theta} \ln t + \rho_t, \tag{5}$$

где невязки ρ_t ограничены и

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^n} \left\|\theta\right\|^2, \theta^{\mathrm{T}} \mathrm{x}_i \ge 1,$$

откуда следует, что скорость сходимости в направлении гиперплоскости, максимизирующей зазор

$$\left\| \frac{\theta_t}{\left\|\theta_t\right\|} - \frac{\hat{\theta}}{\left\|\hat{\theta}\right\|} \right\| = O\left(\frac{1}{\ln t}\right). \tag{6}$$

При этом

$$\min_i \frac{\theta_t^{\mathrm{T}} \mathrm{x}_i}{\left\|\theta_t\right\|} = d - O\left(\frac{1}{\ln t}\right),$$

где

$$d = \max_{\theta} \min_i \frac{\theta^{\mathrm{T}} \mathrm{x}_i}{\left\|\theta\right\|} = \frac{1}{\hat{\theta}} -$$

максимальный зазор.

Как видно из постановки задачи, норма весов не минимизируется, т.е. $\|\theta_t\| \to \infty$, тогда как $\forall i: \theta_t^{\mathrm{T}} \mathrm{x}_i > 0$ при $t \to \infty$, что гарантирует $l'(\theta_t^{\mathrm{T}} \mathrm{x}_i) \to 0$, $L(\theta) \to 0$, а следовательно, сходимость к глобальному минимуму. Соотношение (6) представляет скорость сходимости зазора к максимальному, но не θ_t к $\hat{\theta}$. Таким образом, оценка скорости сходимости предполагает анализ не нормы весов, а лишь направления, т.е. величины $\theta_t / \|\theta_t\|$.

2. Расширенная функция потерь

Для повышения скорости сходимости (6) предложим расширенную функцию потерь вида:

$$l(t;a,b,r,q) = a + \frac{b-a}{\left(1-\left(1-\left(\frac{b-a}{P_0+a}\right)^q\right)\exp\left(-rt\right)\right)^{\frac{1}{q}}}, \tag{7}$$

для которой

$$l'(t;a,b,q)=\frac{r}{q}\left(l(t;a,b,q)-a\right)\times$$
$$\times\left(\left(\frac{l(t;a,b,r,q)-a}{b-a}\right)^{q}-1\right),$$

где нижняя a и верхняя b асимптоты удовлетворяют $0\le a\le b\le 1$, начальное значение нижней асимптоты P_0 удовлетворяет $0\le P_0\le b-2a$, скорость роста функции $r>0$, обобщающий параметр $q>0$, позволяющий регулировать разницу темпов ускорения и замедления роста функции (7). В работах [18, 19] приводится интерпретация предложенной функции потерь через модели динамики популяций в контексте трансдисциплинарных исследований.

Применяя тождество [17]

$$\ln(x)=\lim_{q\to\infty}\frac{1}{q}\left(x^{q}-1\right) \tag{8}$$

к выражению (7), получим следующее определение:

$$l(t;a,b,r)=a+(b-a)\exp\left(c(a,b)\exp(-rt)\right),$$

такое, что $\forall t\in\mathbb{R}: l(t;a,b,r)>0$,

$$\lim_{t\to\infty}l(t;a,b,r)=b-a,\ \lim_{t\to-\infty}l(t;a,b,r)=a,$$
$$c(a,b)=\ln\left(\frac{P_0+a}{b-a}\right),\ c(a,b)<0.$$

Тогда

$$l'(t;a,b,r)=(b-a)c(a,b)r\times$$
$$\times\exp\left(-rt-c(a,b)\exp(-rt)\right), \tag{9}$$

такое, что $\forall t\in\mathbb{R}: l'(t;a,b,r)<0$,

$$\lim_{t\to\infty}l'(t;a,b,r)=\lim_{t\to-\infty}l'(t;a,b,r)=0.$$

Если $q=1$ и $P_0=(b-3a)/2$, то выражение (7) примет вид:

$$l(t;a,b,r)=a+\frac{b-a}{1+\exp(-rt)}. \tag{10}$$

Заметим, что при $a=0$, $b=1$ и $r=1$ функция (7) сводится к сигмоидальной, которая является симметричной с темпом ускорения, эквивалентным темпу замедления.

3. Анализ скорости сходимости

Опустив член $(b-a)c(a,b)r<0$, который задаёт знак, представим выражение (7) в виде:

$$l'(t;c(a,b),r)=-\exp\left(-f(t;c(a,b),r)\right), \tag{11}$$

где

$$f(t;c(a,b),r)=rt-c(a,b)\exp(-rt). \tag{12}$$

Функция $f'(t;a,b,r)$ является строго возрастающей при

$$|t^{*}|>\frac{\ln\left(-c(a,b)\right)}{r}.$$

Для $f(t^{*};c(a,b),r)>0$ функция является положительной при $c(a,b)>-\exp(-1)$. Если $c(a,b)<-1$, то $t^{*}>0$, и, если $c(a,b)<-\exp(r)$, то $t^{*}>1$.

Метод градиентного спуска (4) с учётом градиента (6), построенного на основе расширенной функции потерь (5), примет вид:

$$\theta_{j+1}=\theta_j-\eta\sum_{i=1}^{m}l'\left(\theta_j^{\mathrm{T}}x_i;a,b,r\right)x_i. \tag{13}$$

Проведём анализ скорости сходимости и покажем, что введение расширенной функции потерь в (9) позволяет получить скорость сходимости на (почти) линейно разделимой выборке, эквивалентную:

$$\left\|\frac{\theta_t}{\|\theta_t\|}-\frac{\hat{\theta}}{\|\hat{\theta}\|}\right\|=O\left(\frac{r}{\ln t+W_0\left(c(a,b)/t\right)}\right), \tag{14}$$

где W_0 задаёт W-функцию Ламберта.

Ниже кратко представлены основные положения теоретического обоснования справедливости (14), аналогичные обоснованию, представленному в [4] для выражения (6). Данные положения включают:

a) анализ сходимости (13) к гиперплоскости с максимальным зазором;
b) оценку скорости сходимости зазора к максимальному.

1) Пусть $l'(t;a,b,r)$ задана согласно (12) и $\forall i:\theta_t^{\mathrm{T}}x_i\to\infty$. Если

$$\lim_{t\to\infty}\frac{\theta_t^{\mathrm{T}}x_i}{\|\theta_t\|}=\theta_\infty,\ \text{то}$$
$$\theta_t=\hat{\theta}g_t\left(c(a,b),r\right)+\rho_t\left(c(a,b),r\right), \tag{15}$$

где $g_t(c(a,b),r)\to\infty$, $\forall i:\theta_\infty^{\mathrm{T}}x_i>0$ и

$$\lim_{t\to\infty}\frac{\left\|\rho_t\left(c(a,b),r\right)\right\|}{g_t\left(c(a,b),r\right)}=0.$$

Для компактности представления введём следующие обозначения: $g_t\equiv g_t(c(a,b),r)$, $\rho_t\equiv\rho_t(c(a,b),r)$. Представим градиент расширенной эмпирической функции потерь с учётом (15) в виде:

$$-\nabla L\left(\theta_t,l'(t;a,b,r)\right)=$$
$$=\sum_{i=1}^{m}\exp\left(-f\left(\theta_t^{\mathrm{T}}x_i;c(a,b),r\right)\right)x_i= \tag{16}$$
$$=\sum_{i=1}^{m}\exp\left(-f\left(g_t\theta_\infty^{\mathrm{T}}x_i+\rho_t^{\mathrm{T}}x_i;c(a,b),r\right)\right)x_i.$$

Функция $f(t;a,b,r)$ является возрастающей. Таким образом, при $g_t \to \infty$ выражение $\exp(-f(g_t\,\theta_\infty{}^{\mathrm{T}} x_i + \rho_t{}^{\mathrm{T}} x_i; c(a,b), r))$ становится более отрицательным, так как $\forall i : \theta_\infty{}^{\mathrm{T}} x_i > 0$ и $\|\rho_t\| = o(g_t)$. Следовательно, при условии, что $f(t; c(a,b), r)$ растёт достаточно быстро, наблюдения с минимальным зазором $\arg\min_i \theta_\infty{}^{\mathrm{T}} x_i$ будут формировать сумму (16). Как результат, θ_t, а следовательно, и

$$\hat{\theta} = \frac{\theta_t}{\min_i \theta_\infty{}^{\mathrm{T}} x_i}$$

являются неотрицательной комбинацией опорных векторов [4], описывающих условия Каруша–Куна–Таккера для метода опорных векторов.

$$\hat{\theta} = \sum_{i=1}^{m} \alpha_i x_i ,$$

$$\forall i : \left(\alpha_i \geq 0, \hat{\theta}^{\mathrm{T}} x_i = 1\right) \vee \left(\alpha_i = 0, \hat{\theta}^{\mathrm{T}} x_i > 1\right).$$

Таким образом, θ_∞ пропорционально $\hat{\theta}$.

2) Запишем

$$\theta_t' = -\nabla L\left(\theta_t, l'\left(t;a,b,r\right)\right) =$$

$$= \sum_{i=1}^{m} \exp\left(-f\left(\theta_t{}^{\mathrm{T}} x_i; c(a,b), r\right)\right) x_i .$$

Определим множество индексов

$$S = \arg\min_i \hat{\theta}^T x_i ,$$

таких, что

$$\forall i \in S : \hat{\theta}^{\mathrm{T}} \mathrm{x}_i = 1 .$$

Если $f(t; c(a,b), r)$ растёт достаточно быстро, то при $t \to \infty$ вклад неопорных векторов в формирование градиента становится незначительным:

$$\theta_t' \approx \sum_{i \in S}^{m} \exp\left(-f\left(\theta_t{}^{\mathrm{T}} x_i; c(a,b), r\right)\right) x_i . \qquad (17)$$

Предположим, что ρ_t сходится в направлении a с вектором, ортогональным опорным векторам b. Тогда асимптотическое соотношение (15) примет вид:

$$\theta_t = \hat{\theta} g_t + a h_t + b , \qquad (18)$$

где $h_t = o(g_t)$.

С учётом (18), выражение (17) может быть преобразовано к виду:

$$\hat{\theta} g_t' + \mathrm{a} h_t' \approx$$

$$\approx \sum_{i \in S}^{m} \exp\left(-f\left(\theta_t{}^{\mathrm{T}} x_i g_t + \mathrm{a}^{\mathrm{T}} \mathrm{x}_i h_t'; c(a,b), r\right)\right) x_i .$$

Перепишем последнее соотношение с учётом разложения в ряд Тейлора,

$h_t = o(g_t)$ и $\hat{\theta}^T x_i = 1, \forall i \in S$:

$$\hat{\theta} g_t' \approx \sum_{i \in S}^{m} \exp\left(-f\left(g_t; c(a,b), r\right) + \right.$$

$$\left. + a^T x_i h_t f'\left(g_t; c(a,b), r\right)\right) x_i \approx$$

$$\approx \exp\left(-f\left(g_t; c(a,b), r\right)\right) \times$$

$$\times \sum_{i \in S}^{m} \exp\left(-a^T x_i h_t f'\left(g_t; c(a,b), r\right)\right) x_i .$$

Положим

$$g_t' = \exp\left(-f\left(g_t; c(a,b), r\right)\right),$$

$$h_t' = \left(f'\left(g_t; c(a,b), r\right)\right)^{-1}$$

и $\forall i \in S : \exp(-a^T x_i) = \alpha_i$.

Для решения уравнения относительно q_t воспользуемся $\ln(f'(t)) = o(f(t))$ [4], откуда

$$g_t' = \exp\left(-f\left(g_t; c(a,b), r\right)\right) - \ln\left(f'\left(g_t; c(a,b), r\right)\right),$$

$$g_t = f^{-1}\left(\ln\left(t + C\right)\right).$$

Находя обратную функцию с учётом определения (12)

$$f^{-1}(x) = \frac{x + W_0\left(c(a,b)\exp(-x)\right)}{r},$$

где W_0 задаёт W-функцию Ламберта, получим:

$$g_t = \frac{\ln t + W_0\left(c(a,b)/t\right)}{r}, \qquad (19)$$

что позволяет гарантировать скорость сходимости (14).

Проведём анализ скорости сходимости $g_t(c(a,b), r)$ при различных значениях гиперпараметров $c(a,b)$ и r. Оценим, как введение параметров функции $g_t(c(a,b), r)$ влияет на её скорость роста в сравнении с имеющейся скоростью $\ln t$ и требуемыми скоростями $\sqrt{t}$ и t.

На рис. 1 представлены кривые $1/g_t$, пересекающие кривую $1/g(t)$, где $g(t)$ задаёт требуемую скорость в точках t^*, являющихся решением $s(t^*; c(a,b), r) = 0$, где

$$s\left(t; c(a,b), r\right) = g_t - g\left(t\right),$$

$$\lim_{t \to t^*} g\left(t; c(a,b), r\right) = \frac{g(t)}{r}.$$

Заметим, что пересечение с кривой $1/\ln t$ на обоих рисунках отсутствует, что говорит о более высокой скорости сходимости для предложенной функции потерь. На рис. 1*а* можно видеть, что на начальном участке диаграммы $1/q_t(3, r)$ сходится к $1/\sqrt{t}$, тогда как на рис. 1*б* наилучшие результаты достигаются для скорости $1/q_t(1, r)$, которая сходится к $1/t$. Из рис. 1 также видно различие во влиянии каждого из

гиперпараметров: величина $c(a,b)$ изменяет наклон функций роста $1/g_t(c(a,b),r)$, тогда как параметр r задаёт смещение относительно $1/\sqrt{t}$ и $1/t$. Точные решения t^* для различных $c(a,b)$ и r представлены в табл. 1. Прочерки в таблице указывают на отсутствие пересечений между $1/g_t(c(a,b),r)$ и $1/g(t)$, а следовательно, отсутствие их асимптотической сходимости. Значения для $1/\ln t$ не представлены в таблице ввиду отсутствия пересечений с каждой из анализируемых кривых, что указывает на достоинства предложенного в работе подхода к повышению скорости сходимости.

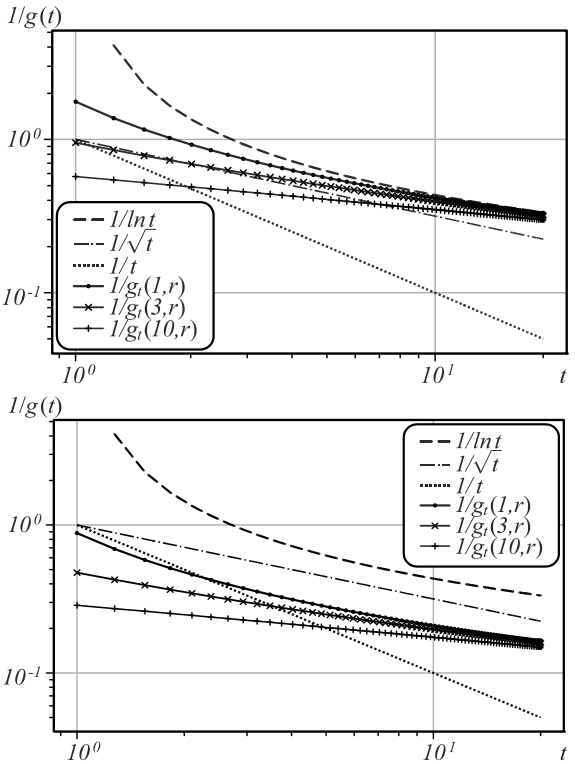

Рис. 1. Анализ функций роста $1/q_t$ в сравнении с исходной $1/\ln t$ и требуемыми $1/q(t)$ при $c(a, b) = \{1; 3; 10\}$: $r = 1$ (а), $r = 0{,}5$ (б)

Табл. 1. Значения решений t^* для заданных $c(a, b) = \{1; 3; 10\}$ и $r = \{1; 0{,}5\}$

$t^*(r;c)$	$r=1$		$r=0{,}5$	
	$1/\sqrt{t}$	$1/t$	$1/\sqrt{t}$	$1/t$
$1/g_r(t;1)$	–	–	0,841	2,298
$1/g_r(t;3)$	2,074	1,094	0,087	3,551
$1/g_r(t;10)$	7,054	2,025	4,714	4,903

Результаты проведённого анализа сходимости представлены для выбранных значений параметров r и $c(a,b)$, задающих область определения расширенной функции потерь. В свою очередь, метод градиентного спуска (8) предполагает оптимизацию данных гиперпараметров на каждой итерации j, что даёт лучшую аппроксимацию функции скорости $1/g(t)$ с помощью $1/g_t(c(a,b),r)$ и приводит к более высокой скорости сходимости, близкой к $O(1/\sqrt{t})$ и $O(1/t)$, гарантируя границу погрешности метода градиентного спуска.

4. Вычислительные эксперименты

Выше описан общий подход к повышению скорости сходимости градиентного метода с помощью расширенной функции потерь. Данный подход может быть адаптирован на случай более широкого класса градиентных методов при решении задач классификации с целью минимизации вычислительных затрат.

Для наблюдения эффекта неявной регуляризации вычислительные эксперименты предполагали многоклассовую классификацию изображений на наборах MNIST и Fashion MNIST с помощью модели нейронной сети как наиболее широкое приложение глубокого обучения [20]. Модель включала два скрытых слоя с 10 нейронами на каждом из них. Для обучения сети использовался метод стохастического градиента, построенный на основе (13) [19]. Функции активации на скрытых и выходном слоях были построены на основе расширенной функции потерь (10) с оптимальными гиперпараметрами r, a и b.

Целью вычислительной экспериментов являлся анализ влияния «неявного» смещения, задаваемого методом оптимизации, на выход последнего слоя сети в случае, когда последний скрытый слой становится почти линейно разделимым после заданного количества итераций [4, 7, 21]. Обучающие выборки для каждого набора данных были разбиты на подвыборки для обучения и контроля на основе 5-fold CV. Обучение сети выполнялось при $n_{итераций} = 1000$ и $n_{пакетов} = 25$ для реализации метода стохастического градиента, оптимизация гиперпараметров – с помощью случайного поиска [22] со случайным выбором 15 % возможных сочетаний параметров. Для снижения временных затрат оптимизация параметров r, a и b выполнялась на $n_{итераций} = 1$, что должно оказать влияние на результат классификации уже на начальном этапе обучения. На рис. 2 и 3 приведены результаты обучения модели на наборах MNIST и Fashion MNIST.

Прежде всего, на представленных рисунках можно наблюдать эффект неявной регуляризации: кривые функции потерь на обучении приближаются к нулю, кривые функции потерь на контроле начали возрастать, но точности классификации на контрольных выборках по-прежнему растут с каждой итерацией.

Из рис. 2 можно видеть, что рост кривой функции потерь на всем интервале обучения от 1 до 1000 итераций для стандартной модели $(c(a,b),r)_{def}$ в сравнении с расширенной моделью $(c(a,b),r)_{opt}$ выше в 88,3 раза (см. рис. 2*а*). Прирост точности при использовании расширенной модели для 1, 100 и 1000 итераций на 0,54 %, 0,53 % и 0,36 %, соответственно (рис. 2*в*). Данный результат указывает на то, что при использовании расширенной модели достаточно меньшего количества итераций для обеспечения приемлемой точ-

ности классификации без существенного прироста, что существенным образом снижает вычислительные результаты.

В случае набора Fashion MNIST (см. рис. 3) рост кривой функции потерь на всем интервале для стандартной модели в сравнении с расширенной моделью выше в 3,4 раза (см. рис. 3*а*). Прирост точности для 1, 100 и 1000 итераций достигает 30,8 %, 2,94 % и 1,05 %, соответственно (см. рис. 3*в*). Данный результат также указывает на достоинства расширенной модели, требующей меньшего количества итераций для значительного прироста точности классификации, что также приводит к снижению вычислительных результатов.

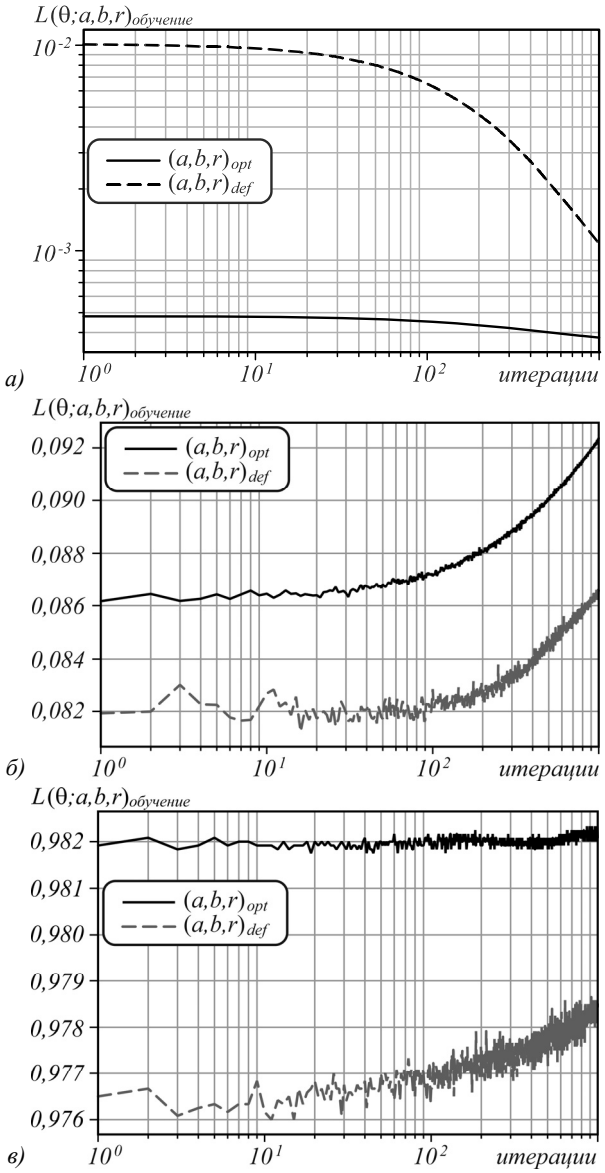

Рис. 2. Результат обучения модели на наборе MNIST: функция потерь на обучении (а); функция потерь на контроле (б), точность классификации на контроле (в)

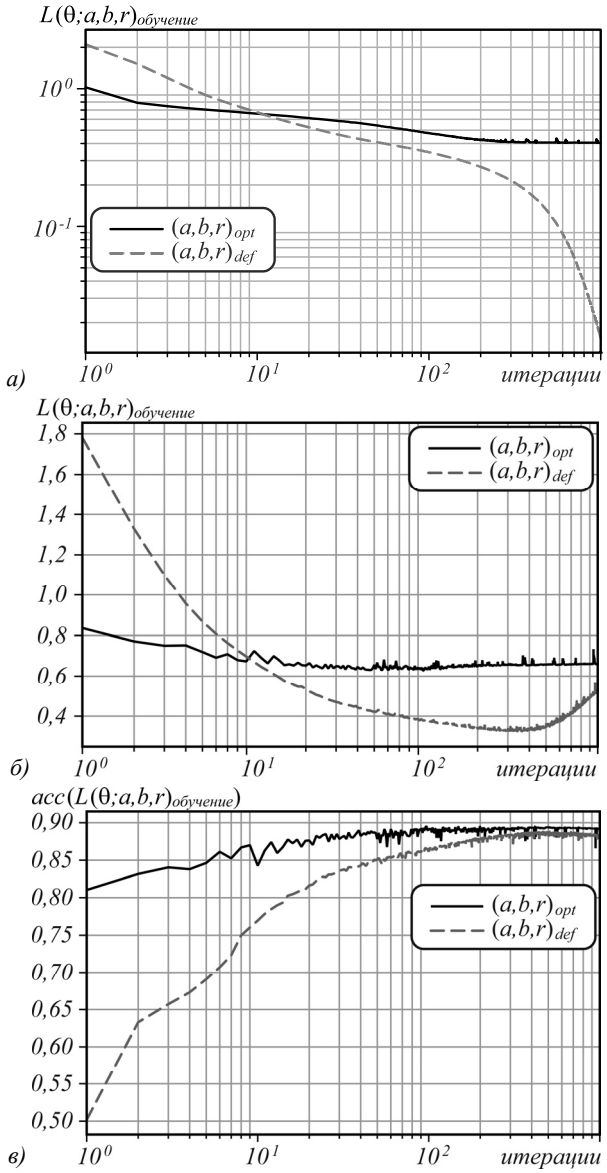

Рис. 3. Результат обучения модели на наборе Fashion MNIST: функция потерь на обучении (а), функция потерь на контроле (б), точность классификации на контроле (в)

Заключение

В данной работе был предложен способ снижения вычислительных затрат в глубоком обучении с помощью расширенной функции потерь с гиперпараметрами. Теоретический анализ показал, что обучение модели с расширенной функцией потерь приводит к более высокой скорости сходимости, близкой к $O(1/\sqrt{t})$ и $O(1/t)$, гарантируя границу погрешности метода градиентного спуска. Эмпирический анализ выявил, что при построении модели обучения на основе расширенной функции потерь достаточно меньшего количества итераций как для получения устойчивого приемлемого значения точности классификации без существенного прироста, так и для обеспечения значительного прироста данного значения, что существенным образом снижает вычислительные результаты.

Дальнейшие направления исследований связаны с более детальным теоретическим анализом и построением асимптотических оценок скоростей сходимости на различных интервалах определения гиперпараметров.

Благодарности

Автор выражает признательность рецензенту за замечания и рекомендации, которые привели к повышению качества представления материалов исследования. Работа выполнена при поддержке грантов Президента РФ (проект № МК-6218.2018.9), Минобрнауки РФ (проект № 074-U01), РФФИ (№ 18-37-00219), а также проекта DATACROSS Центра Превосходства, финансируемого Правительством Хорватии и Европейским Союзом через Европейский фонд регионального развития – Операционная программа конкурентоспособности и сплочения (KK.01.1.1.01.0009).

Литература

1. **LeCun, Y.** Deep learning / Y. LeCun, Y. Bengio, G. Hinton // Nature. – 2015. – Vol. 521(7553). – P. 436-444. – DOI: 10.1038/nature14539.
2. **Goodfellow, I.** Deep learning / I. Goodfellow, Y. Bengio, A. Courville. – Cambridge, London: The MIT Press, 2016. – 800 p. – ISBN: 978-0-262-03561-3.
3. **Neyshabur, B.** In search of the real inductive bias: On the role of implicit regularization in deep learning [Electronical Resource] / B. Neyshabur, R. Tomioka, N. Srebro // arXiv preprint. – URL: https://arxiv.org/abs/1412.6614 (request date 5.12.2019).
4. **Soudry, D.** The implicit bias of gradient descent on separable data / D. Soudry, E. Hoffer, M.S. Nacson, S. Gunasekar, N. Srebro // Journal of Machine Learning Research. – 2018. – Vol. 19. – P. 1-57.
5. **Zhang, C.** Understanding deep learning requires rethinking generalization / C. Zhang, S. Bengio, M. Recht, O. Vinyals // arXiv preprint arXiv:1611.03530v2, 2017.
6. **Hoffer, E.** Train longer, generalize better: closing the generalization gap in large batch training of neural networks [Electronical Resource] / E. Hoffer, I. Hubara, D. Soudry // arXiv preprint. – URL: https://arxiv.org/abs/1705.08741 (request date 5.12.2019).
7. **Nacson, M.S.** Convergence of gradient descent on separable data / M.S. Nacson, J. Lee, S. Gunasekar, N. Srebro, D. Soudry // 2019 22nd International Conference on Artificial Intelligence and Statistics (AISTATS). – 2019. – Vol. PMLR 89. – P. 3420-3428.
8. **Gunasekar, S.** Characterizing implicit bias in terms of optimization geometry / S. Gunasekar, J. Lee, D. Soudry, N. Srebro // 2018 35th International Conference on Machine Learning (ICML). – 2018. – Vol. PMLR 80. – P. 1832-1841.
9. **Ma, C.** Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion / C. Ma, K. Wang, Y. Chi, Y. Chen // 2018 35th International Conference on Machine Learning (ICML). – 2018. – Vol. PMLR 80. – P. 3345-3354.
10. **Kingma, D.P.** Adam: A method for stochastic optimization [Electronical Resource] / D.P. Kingma, J.L. Ba // arXiv preprint. – URL: https://arxiv.org/abs/1412.6980 (request date 5.12.2019).
11. **Duchi, J.** Adaptive subgradient methods for online learning and stochastic optimization / J. Duchi, E. Hazan, Y. Singer // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2121-2159.
12. **Zeiler, M.D.** ADADELTA: An adaptive learning rate method [Electronical Resource] / M.D. Zeiler // arXiv preprint. –URL: https://arxiv.org/abs/1212.5701 (request date 5.12.2019).
13. **Kim, H.S.** Convergence analysis of optimization algorithms [Electronical Resource] / H.S. Kim, J.H. Kang, W.M. Park, S.H. Ko, Y.H. Cho, D.S. Yu, Y.S. Song, J.W. Choi // arXiv preprint. – URL: https://arxiv.org/abs/1707.01647 (request date 5.12.2019).
14. **Ruder, S.** An overview of gradient descent optimization algorithms [Electronical Resource] / S. Ruder // arXiv preprint. – URL: https://arxiv.org/abs/1609.04747 (request date 5.12.2019).
15. **Wilson, A.C.** The marginal value of adaptive gradient methods in machine learning / A.C. Wilson, R. Roelofs, M. Stern, N. Srebro, B. Recht // 2017 31st Conference on Neural Information Processing Systems (NIPS). – 2017. – P. 1-11.
16. **Воронцов, К.В.** Математические методы обучения по прецедентам (теория обучения машин) [Электронный ресурс] / К.В. Воронцов. – URL: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (дата обращения 01.12.19).
17. **Castañeda, A.R.S.** New formulation of the Gompertz equation to describe the kinetics of untreated tumors / A.R.S. Castañeda, E.R. Torres, N.A.V. Goris, [et al.] // PLoS ONE. – 2019. – Vol. 14, Issue 11. – e0224978.
18. **Kulikovskikh, I.** BioGD: Bio-inspired robust gradient descent / I. Kulikovskikh, S. Prokhorov, T. Lipić, T. Legović, T. Šmuc // PLoS ONE. – 2019. – Vol. 14, Issue 7. – e0219004.
19. **Kulikovskikh, I.** An SGD-based meta-learner with "growing" descent / I. Kulikovskikh, S. Prokhorov, T. Legović, T. Šmuc // Journal of Physics: Conference Series. – 2019. – Vol. 1368. – 052008.
20. **Савченко, А.В.** Метод максимально правдоподобных рассогласований в задаче распознавания изображений на основе глубоких нейронных сетей / А.В. Савченко // Компьютерная оптика. – 2017. – Т. 41, № 3. – С. 422-430. – DOI: 10.18287/2412-6179-2017- 41-3-422-430.
21. **An, S.** How can deep rectifier networks achieve linear separability and preserve distances? / S. An, F. Boussaid, M. Bennamoun // 2015 32nd International Conference on Machine Learning (ICML). – 2015. – Vol. PMLR 375. – P. 514-523.
22. **Bergstra, J.** Random search for hyperparameter optimization / J. Bergstra, Y. Bengio // Journal of Machine Learning Research. – 2012. – Vol. 13. – P. 281-305.

Сведения об авторе

Куликовских Илона Марковна является постдокторским исследователем на факультете электротехники и вычислительной техники Загребского университета и в Лаборатории машинного обучения и представления знаний в Институте Руджер Бошкович. Работает доцентом на кафедре информационных систем и технологий

Самарского университета. В 2008 году окончила Самарский государственный аэрокосмический университет по специальности «Автоматизированные системы обработки информации и управления». В 2011 году защитила диссертацию на соискание степени кандидата наук по специальности «Математическое моделирование, численные методы и комплексы программ» в Самарском национальном исследовательском университете. Имеет более 100 публикаций, среди которых 6 книг и учебных пособий. Область научных интересов: машинное обучение, анализ сигналов, статистический анализ данных, вычисления на основе принципов организации живых организмов и непрерывное обучение. E-mail: *kulikovskikh.i@gmail.com* .

ГРНТИ: 28.23.25
Поступила в редакцию 13 октября 2019 г. Окончательный вариант – 13 декабря 2019 г.

Reducing computational costs in deep learning on almost linearly separable training data

I.M. Kulikovsvkikh [1,2,3]
[1] *Samara National Research University,*
443086, Russia, Samara, Moskovskoe Shosse 34,
[2] *Faculty of Electrical Engineering and Computing, University of Zagreb,*
10000, Croatia, Zagreb, Unska 3,
[3] *Rudjer Boskovic Institute,*
10000, Croatia, Zagreb, Bijenicka cesta 54

Abstract

Previous research in deep learning indicates that iterations of the gradient descent, over separable data converge toward the L2 maximum margin solution. Even in the absence of explicit regularization, the decision boundary still changes even if the classification error on training is equal to zero. This feature of the so-called "implicit regularization" allows gradient methods to use more aggressive learning rates that result in substantial computational savings. However, even if the gradient descent method generalizes well, going toward the optimal solution, the rate of convergence to this solution is much slower than the rate of convergence of a loss function itself with a fixed step size. The present study puts forward the generalized logistic loss function that involves the optimization of hyperparameters, which results in a faster convergence rate while keeping the same regret bound as the gradient descent method. The results of computational experiments on MNIST and Fashion MNIST benchmark datasets for image classification proved the viability of the proposed approach to reducing computational costs and outlined directions for future research.

Keywords: implicit regularization, gradient method, convergence rate, linear separability, image classification.

Citation: Kulikovskikh IM. Reducing computational costs in deep learning on almost linearly separable training data. Computer Optics 2020; 44(2): 282-289. DOI: 10.18287/2412-6179-CO-645.

Acknowledgements: This work was supported by the Russian Federation President's grant (Project No. MK-6218.2018.9), the Ministry of Education and Science of the Russian Federation (Project No. 074-U01), RFBR (Project No. 18-37-00219), and the 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).

References

[1] LeCun Y, Bengio Y. Deep learning. Nature 2015; 521(7553): 436-444. DOI: 10.1038/nature14539.

[2] Goodfellow I, Bengio Y, Courville Y. Deep learning. Cambridge, London: The MIT Press; 2016. ISBN: 978-0-262-03561-3.

[3] Neyshabur B, Tomioka R, Srebro N. In search of the real inductive bias: On the role of implicit regularization in deep learning. Source: ⟨https://arxiv.org/abs/1412.6614⟩.

[4] Soudry D, Hoffer E, Nacson MS, Gunasekar S, Srebro N. The implicit bias of gradient descent on separable data. J Mach Learn Res 2018; 19: 1-57.

[5] Zhang C, Bengio S, Recht M, Vinyals O. Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530v2, 2017.

[6] Hoffer E, Hubara I, Soudry D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. Source: ⟨https://arxiv.org/abs/1705.08741⟩.

[7] Nacson MS, Lee J, Gunasekar S, Srebro N, Soudry D. Convergence of gradient descent on separable data. 2019 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019; PMLR 89: 3420-3428.

[8] Gunasekar S, Lee J, Soudry D, Srebro N. Characterizing implicit bias in terms of optimization geometry. 2018 35th International Conference on Machine Learning (ICML) 2018; PMLR 80: 1832-1841.

[9] Ma C, Wang K, Chi Y, Chen Y. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. 2018 35th International Conference on Machine Learning (ICML) 2018; PMLR 80: 3345-3354.

[10] Kingma DP, Ba JL. Adam: A method for stochastic optimization. Source: ⟨https://arxiv.org/abs/1412.6980⟩.

[11] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res 2011; 12: 2121-2159.

[12] Zeiler MD. ADADELTA: An adaptive learning rate method. Source: ⟨https://arxiv.org/abs/1212.5701⟩.

[13] Kim HS, Kang JH, Park WM, Ko SH, Cho YH, Yu DS, Song YS, Choi JW. Convergence analysis of optimization algorithms. Source: ⟨https://arxiv.org/abs/1707.01647⟩.

[14] Ruder S. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747, 2016.

[15] Wilson AC, Roelofs R, Stern M, Srebro N, Recht B. The marginal value of adaptive gradient methods in machine learning. 2017 31st Conference on Neural Information Processing Systems (NIPS) 2017: 1-11.

[16] Vorontsov KV. Mathematical methods for supervised learning (machine learning theory) [In Russian]. Source:

⟨http:// www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf⟩.
[17] Castaneda ARS, Torres ER, Goris NAV, González MM, Reyes JB, González VGS, et al. New formulation of the Gompertz equation to describe the kinetics of untreated tumors. PLoS ONE 2019; 14(11): e0224978.
[18] Kulikovskikh I, Prokhorov S, Lipić T, Legović T, Šmuc T. BioGD: Bio-inspired robust gradient descent. PLoS ONE 2019; 14(7): e0219004.
[19] Kulikovskikh I, Prokhorov S, Legović T, Šmuc T. An SGD-based meta-learner with "growing" descent. J Phys: Conf Ser 2019; 1368: 052008.
[20] Savchenko AV. Maximum-likelihood dissimilarities in image recognition with deep neural networks. Computer Optics 2017; 41(3): 422-430. DOI: 10.18287/2412-6179-2017-41-3-422-430.
[21] An S, Boussaid F, Bennamoun M. How can deep rectifier networks achieve linear separability and preserve distances? 2015 32nd International Conference on Machine Learning (ICML) 2015; PMLR 375: 514-523.
[22] Bergstra J, Bengio Y. Random search for hyperparameter optimization. J Mach Learn Res 2012; 13: 281-305.

Author's information

Ilona M. Kulikovskikh is a postdoctoral researcher of Electrical Engineering and Computing faculty at the University of Zagreb and the Laboratory for Machine Learning and Knowledge Representation at the Ruđer Bošković Institute. She is an associate professor of Information Systems and Technologies department at Samara National Research University. She defended her graduation work in Computer Science at Samara State Aerospace University with distinction in 2008 and received her PhD in Signal Processing, Data Processing and Automation Control from Samara National Research University in 2011. She is an author of more than 100 refereed scientific papers published in Russian and in English. Among them are six co-authored monographs and study books. Her research interests are in the areas of machine learning, signal processing, statistical data processing, bio-inspired computing, and life-long learning.
E-mail: *kulikovskikh.i@gmail.com* .

Received October 13, 2019. The final version – December 13, 2019.