A method of implicit regularization based on the phenomena of retrieval-induced forgetting (RIF)

I.M. Kulikovskikh[1], S.A. Prokhorov[1]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086

Abstract. Deep learning models have been successfully applied to a variety of real-world problems due to its ability to recognize a complex structure in large datasets through revealing non-trivial relationships among multiple levels of data representations. However, widely used in deep learning gradient-based algorithms may cause numerous difficulties on account of limited memory. While recent studies addressed the problem of the lack of an external memory, and, thus, improved the generalization ability, the proposed solutions introduced a kind of implicit regularization which seems poorly controlled and, as a consequence, decrease the interpretability of learning models. In an attempt to deepen understanding the nature of generalization ability, the present study is aimed at looking at implicit regularization from a psychological perspective. This research puts forward a method of implicit regularization based on the phenomena of retrieval-induced forgetting (RIF). The findings of this study may greatly assist in solving the major problems of proper understanding the deep learning procedure, improving the generalization ability, and the capacity control.

Keywords: Implicit regularization, Guessing technique, Retrieval-induced forgetting.

1. Introduction

Deep learning allows highly efficient models that have improved the state-of-the-art in a broad range of applications. This success may be attributed to the ability to recognize a complex structure in large datasets through building the relationship between the multiple layers of learning models and the multiple levels of data representations [1-3]. However, there seems a problem of distinguishing between deep learning models that have different generalization performance. The traditional approaches like VC dimension, Rademacher complexity fail to explain why these models may generalize well in practice [4].

Theory suggests different forms of explicit regularization to ensure small generalization error as is the case with a large number of model parameters. However, regularization may be also introduced by modifying the optimization method through drop-outs, weight decays, gradient noise and etc. A review of the literature on this issue indicates that some sort of implicit regularization [5-11] may be essential in a proper generalization of deep learning models. This type of regularization tries to find a solution with small complexity, but neither does not include a penalty term nor does not directly modify the optimization procedure. According to [8] the generalization ability is controlled by the geometry of the model parameter space and the empirical optimization procedure attuned to this geometry.

In an attempt to deepen understanding the nature of generalization ability, the present study is aimed at looking at implicit regularization from a psychological perspective. The present research proposes an implicit regularization method based on the extended logistic regression with respect to forgetting and guessing factors [12]. Following from the definitions of floor and ceiling effects in statistics [13-15] and psychology [16, 17], these factors allow an improved gradient descent with a thinking trace.

2. Problem statement

The presence of floor or/and ceiling effects may cause the negative log-likelihood failure to converge due to the problem of clear separation between the classes [13-15]. The logit function may go to $-\infty$ for 0 successes and ∞ for 0 failures. These effects may be clearly interpreted from the perspective of cognitive and educational psychology [16, 17]. According to the surveyed sources, a ceiling effect occurs when a measure (psychological/intelligence test) has a marked upper limit for responses that mostly concentrate at or near the limit (ceiling). A floor effect, in contrast, occurs when a measure imposes a distinct lower level so that a large concentration of responses is at or near this limit (floor). These effects can be caused by a number of reasons. The most convincing of them is the following: if the measure involves a task with an upper/lower limit, such as a number of correct responses, this task can be found too easy/difficult. As a consequence, the assessment results indicate a nearly perfect/almost zero score on the measure. A lack of variance due to a ceiling or floor effect casts doubt on the validity of the measure and the performance outcomes.

To cope with ceiling and floor effects, psychologists use different approaches. The most obvious is varying a task difficulty by changing a number of potential responses in multiple-choice testing [18-20]. But an increase in the number of distractors may lead to a decrease in proportions of correct responses. Test-takers are likely to acquire false knowledge instead of enhancing retention of the material. As a result, such test format may increase test-takers' exposure to misinformation.

The authors [18, 21, 22], however, stated that multiple-choice testing can stimulate deep learning and increase long-term retention. In accordance with these studies, multiple-choice testing can stimulate the type of retrieval processes known to improve learning. First, retrieval practice can enhance long-term retention of the tested material. Then, it can also impair later recall of the nontested material [23]. This phenomenon, known as **retrieval-induced forgetting (RIF)** [18, 21], was first described in terms of suppressing memories that become not relevant for a given situation. To stimulate these retrieval processes, test-takers should be provided with a metacognitive strategy to encourage more complex thinking. This strategy is aimed at considering all the alternatives to cogitate not only why the selected answer is correct, but also why distractors are incorrect. In addition, test-takers should engage in this metacognitive strategy even if they are certain what answer is correct. Applying metacognitive strategies, in turn, may pose the other serious assessment problem: if test-takers can eliminate some responses based on critical analysis, they can get the correct answer with **guessing**, the level of which is often difficult to assess correctly [24, 25].

Thus, to address the problem of proper assessment in presence of floor and ceiling factors, Macready and Dayton [26] made two underlying assumptions:

1) an observed failure in test results stems from **forgetting**;
2) an observed success in test results is attributed to **guessing**.

Following this, let us pose the problem of implicit regularization that consists in optimizing the thinking traces based on forgetting and guessing factors.

Let $(x_i, y_i)_{i=1}^m$ be independent and identically distributed observations with binary responses $y_i \in \{0,1\}$. The matrix $X \in \mathbf{R}^{m \times n}$ can be viewed either as vectors of predictors $x_i \in \mathbf{R}^n$ or as vectors of features $x^j \in \mathbf{R}^m$. Then, for any vector $\theta \in \mathbf{R}^n$ of regression coefficients logistic regression models the class conditional probabilities $p(x_i, \theta) = P(y_i = 1 | x_i, \theta)$ by $\ln\left(\frac{p(x_i,\theta)}{1-p(x_i,\theta)}\right) = \theta^T x_i$ [13]. Having defined the parameters of ordinary logistic regression, let us denote forgetting and guessing factors as follows.

Let $c^{guess} \equiv c^g$, $c^{forget} \equiv c^f$, and $c^{g,f} \equiv \left(c^g, c^f\right)$, where $c^{g,f} \in [0,1]$. Using the denotations introduced the logit function may be extended with regard to $c^{g,f}$ [12] as

$$g\left(p\left(x_i, \theta, c^{g,f}\right)\right) = \ln\left(\frac{p\left(x_i, \theta, c^{g,f}\right)}{1 - p\left(x_i, \theta, c^{g,f}\right)}\right), \quad (1)$$

where

$$p\left(x_i, \theta, c^{g,f}\right) = c^g + \frac{1 - \left(c^g + c^f\right)}{1 + \mathrm{e}^{-\theta^T x_i}}. \quad (2)$$

Taking into account the definitions (1) and (2), let us pose the following problem.

Problem 1. For each $x_i \in \mathbf{R}^n$, $\theta \in \mathbf{R}^n$, $\{m,n\} \in \mathbf{N}$, and $c^{g,f}_{\{m,n\}} \in [0,1]$,

$$\ln L\left(\theta, c^{g,f}\right)_{\{m,n\}} = -\sum_i y_i \ln\left(p\left(x_i, \theta, c^{g,f}\right)\right) + (1 - y_i)\ln\left(1 - p\left(x_i, \theta, c^{g,f}\right)\right)$$

needs to be minimized to solve the problem

$$\left(\theta, c^{g,f}\right)_{\{m,n\}}{}^* = \underset{\substack{\theta \in \mathbf{R}^n \\ c^{g,f} \in [0,1]}}{\arg\min} \ln L\left(\theta, c^{g,f}\right)_{\{m,n\}}.$$

Problem 1 allows to introduce **a thinking trace** with respect to $\{m,n\} \in \mathbf{N}$ in the form:

$$c^{g,f}_{\{m,n\}}{}^* = \underset{\substack{\theta \in \mathbf{R}^n \\ c^{g,f} \in [0,1]}}{\arg\min} \ln L\left(\theta, c^{g,f}\right)_{\{m,n\}}. \quad (3)$$

To examine the influence of $c^{g,f}_{\{m,n\}}{}^*$ on the convergence issues and generalization ability, the next part of this paper will present us with the experimental evidence on the proposed measure.

3. Experiments

A brief description of 3 datasets used to validate the theoretical results is given in Table 1. The information includes the values of $\{m,n\}$ and the class distribution. These datasets are freely available from UCI Machine Learning repository.

Table 1. A brief description of datasets.

dataset	*m*			*n*
	$y_i \in \{0,1\}$	$y_i = 0$	$y_i = 1$	
Vertebral Column (vertebral)	309	100	209	6
Liver Disorder (liver)	345	145	200	8
Pima Indians Diabetes (pima)	768	500	268	9

It is known that small and unbalanced datasets are the most obvious reason for floor and ceiling effects. Thus, the chosen datasets present a different combination of the number of observations *m* and the number of features *n* to increase the chance of identifying these effects. In addition, the design of experiments suggested varying the number of observations $m \in \{ak + b \mid k \in [0, K]\}$, where the number of steps *K=4*. The rate of thinking strategy was calculated subject to *a=15*, *b=20*.

The datasets were divided into the training subset and the validation subset using 3-fold cross validation. As for small to moderate sample sizes the resampling estimates are better than the asymptotic estimates, the bootstrap method was adopted to provide reliable results.

To report an improvement in guessing and forgetting within each trial the following indicators were designed:

$$\delta_{\{m,n\}} = \frac{\ln L(\theta)_{\{m,n\}} - \ln L\left(\theta, c^{g,f}\right)_{\{m,n\}}}{\ln L\left(\theta, c^{g,f}\right)_{\{m,n\}}}$$

and

$$\varepsilon_{\{m,n\}} = \frac{\ln L\left(\theta, c^{g,f}\right)_{\{m_{k+1},n\}} - \ln L\left(\theta, c^{g,f}\right)_{\{m_k,n\}}}{\delta_{\{m,n\}}},$$

where the rate of improvement $\varepsilon_{\{m,n\}}$ is set upon a scale

$\varepsilon_{\{m,n\}} \in \{-1, -0.1, -0.01, 0, 0.01, 0.1, 1\}$.

The scale ranges between marked improvement ($\varepsilon_{\{m,n\}} < -0.01$) and a lack of improvement ($0 \le \varepsilon_{\{m,n\}} < 1$) with little improvement ($-0.1 \le \varepsilon_{\{m,n\}} < 0$) in borderline cases.

The experiments were aimed at analyzing $\delta_{\{m,n\}}$ and $\varepsilon_{\{m,n\}}$ based on the cross-validation estimates of prediction error for the ordinary $\ln L(\theta)_{\{m,n\}}$ and extended $\ln L\left(\theta, c^{g,f}\right)_{\{m,n\}}$ loss functions (see Table 2).

Table 2. Cross-validation estimates of prediction error.

	m	$c^{g,f}$	$\ln L\left(\theta, c^{g,f}\right)_{\{m,n\}}$	$\delta_{\{m,n\}}$	$\varepsilon_{\{m,n\}}$
vertebral ($\varepsilon_{\{m,n\}} < -0.01$)	20	(0.058,0.115)	0.4567	0.3593	-
	35	(0.0049,0)	0.6259	0.1933	0.8755
	50	(0.0049,0)	0.6333	0.1534	0.0474
	65	(0.0049,0)	0.6318	0.1402	**-0.0108**
	80	(0.0049,0)	0.6224	0.1226	**-0.0774**
liver ($\varepsilon_{\{m,n\}} < -1$)	20	(0,0.0073)	0.81139	0.07104	-
	35	(0,0.0808)	0.89288	0.03573	2.28065
	50	(0,0.0073)	0.88546	0.02357	**-0.31469**
	65	(0,0.0073)	0.86795	0.02061	**-0.84933**
	80	(0,0.0073)	0.83438	0.01727	**-1.94406**
pima ($\varepsilon_{\{m,n\}} < -1$)	20	(0.0955,0.1176)	2.25079	0.19788	-
	35	(0.1176,0.0882)	2.19594	0.22568	**-0.24301**
	50	(0.1176,0.0955)	2.07351	0.15933	**-0.76844**
	65	(0.0661,0.0906)	1.87494	0.11166	**-1.77834**
	80	(0.0661,0.0906)	1.71529	0.08097	**-1.97179**

The values highlighted in bold correspond to $\varepsilon_{\{m,n\}} < 0$ as these values reflect the ever-increasing rate of improvement in forgetting and guessing $c^{g,f}$. It is also expected that the level of influence of c^g and c^f brings down within each trial as the proposed measure is an attempt to model short-term memory with a guessing technique. To keep the positive trend in $\delta_{\{m,n\}}$, the thinking traces based on (3) need to be extended to the case of long-term memory.

4. Conclusions

The present study explores the nature of implicit regularization from a psychological perspective. Combining the definitions of floor and ceiling effects in statistics and psychology, this research introduces forgetting and guessing factors and, with respect to them, puts forward an extension of logistic regression and the improved machine learning procedure with thinking traces. The in-depth analysis showed that the inclusion of these factors in the model results in improved convergence of log-likelihood and depends on the relationship between forgetting and guessing strategies. Taking into consideration the current state-of-art in deep learning, it seems promising to extend the proposed measure to the case of long-term memory with a guessing strategy.

5. Acknowledgments

This work was supported by the Russian Federation President grant MK-6218.2018.9 and the Ministry of Education and Science of the Russian Federation grant 074-U01.

6. References

[1] LeCun, Y. Deep learning / Y. LeCun, Y. Bengio, G. Hinton // Nature. – 2015. – Vol. 521. – P. 436-444.

[2] Ronen, R. Why & When Deep Learning Works: Looking Inside Deep Learnings / R. Ronen // Access mode: arXiv:1705.03921 (23.06.2017).

[3] Shwartz-Ziv, R. Opening the Black Box of Deep Neural Networks via Information / R. Shwartz-Ziv, N. Tishby // Access mode: arXiv:1706.05394 (23.09.2017).

[4] Zhang, C. Understanding deep learning requires rethinking generalization / C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals // Access mode: arXiv:1611.03530 (23.09.2017).

[5] Fan, Q. Convergence of batch gradient learning with smoothing regularization and adaptive momentum for neural networks / Q. Fan, W. Wu, J.M. Zurada // Springerplus. – 2016. DOI: 10.1186/s40064-016-1931-0.

[6] Gunasekar, S. Implicit Regularization in Matrix Factorization / S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, N. Srebro // Access mode: arXiv:1705.09280 (23.09.2017).

[7] Lin, J. Generalization Properties and Implicit Regularization for Multiple Passes SGM / J. Lin, R. Camoriano, L. Rosasco // The Proceedings of the 33rd International Conference on Machine Learning. – NY, USA, 2016. – P. 2340-2348.

[8] Neyshabur, B. Implicit Regularization in Deep Learning / B. Neyshabur // Access mode: arXiv:1709.01953 (23.09.2017).

[9] Neyshabur, B. In search of the real inductive bias: On the role of implicit regularization in deep learning / B. Neyshabur, R. Tomioka, N. Srebro // Access mode: arXiv: 1412.6614 (23.09.2017).

[10] Neyshabur, B. Geometry of Optimization and Implicit Regularization in Deep Learning // B. Neyshabur, R. Tomioka, R. Salakhutdinov, N. Srebro // Access mode: arXiv:1705.03071 (23.09.2017).

[11] Zhang, C. Understanding deep learning requires rethinking generalization / C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals // Access mode: arXiv:1611.03530 (23.09.2017).

[12] Kulikovskikh, I.M. Cognitive validation map for early occupancy detection in environmental sensing / I.M. Kulikovskikh // Engineering Applications of Artificial Intelligence. – 2017. – Vol. 65. – P. 330-335.

[13] Kulikovskikh, I.M. Minimizing the effects of floor and ceiling to improve the convergence of log-likelihood / I.M. Kulikovskikh, S.A. Prokhorov // Procedia Engineering. – 2017. – Vol. 201. – P. 779-788.

[14] Donnelly, S. Empirical logit analysis is not logistic regression / S. Donnelly, J. Verkuilen // Journal of Memory and Language. – 2017. – Vol. 94. – P. 28-42.

[15] Hastie, T. The elements of statistical learning: Data mining, inference, and prediction (2nd ed.) / T. Hastie, R. Tibshirani, J. Friedman. – Springer Series in Statistics, 2013. – 745 p.

[16] Everitt, B.S. The Cambridge dictionary of statistics / B.S. Everitt. – Cambridge: Cambridge University Press, 2010. – 480 p.

[17] Groth-Marnat, G. Handbook of psychological assessment / G. Groth-Marnat, A.J. Wright. – Wiley, 2016. – 768 p.

[18] Bjork, E.L. Multiple-choice testing as a desired difficulty in the classroom / E.L. Bjork, J.L. Little, B.C. Storm // Journal of Applied Research in Memory and Cognition. – 2014. – Vol. 3(3). – P 165-170.

[19] Elliott, G. Measuring forgetting: A critical review of accelerated long-term forgetting studies / G. Elliott, C.L. Isaac, N. Muhlert // Cortex. – 2014. – Vol. 54. – P. 16-32.

[20] Lesage, E. Scoring methods for multiple choice assessment in higher education - Is it still a matter of number right scoring or negative marking? / E. Lesage, M. Valcke, E. Sabbe // Studies in Educational Evaluation. – 2013. – Vol. 39. – P. 188-193.

[21] Bjork, E.L. Can multiple-choice testing induce desirable difficulties? Evidence from the Laboratory and the Classroom / E.L. Bjork, N.C. Soderstrom, J.L. Little // The American Journal of Psychology. – 2015. – Vol. 128(2). – P. 229-239.

[22] Little, J.L. Optimizing multiple-choice tests as tools for learning / J.L. Little, E.L. Bjork // Memory & Cognition. – 2015. – Vol. 43. – P. 14-26.
[23] Chan, J.C.K. When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing / J.C.K. Chan // Journal of Memory and Language. – 2009. – Vol. 61(2). – P. 153-170.
[24] Kubinger, K.D. On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format / K.D. Kubinger, S. Holocher-Ertl, M. Reif, C. Hohensinn, M. Frebort // International Journal of Selection and Assessment. – 2010. – Vol. 18(1). – P. 111-115.
[25] Kulikovskikh, I.M. Promoting collaborative learning through regulation of guessing in clickers / I.M. Kulikovskikh, S.A. Prokhorov, S.A. Suchkova // Computers in Human Behavior. – 2017. – Vol. 75. – P. 81-91.
[26] Macready, G.B. The use of probabilistic models in the assessment of mastery / G.B. Macready, C.M. Dayton // Journal of Educational Statistics. – 1977. – Vol. 2. – P. 99-120.