Psychological perspectives on implicit regularization: A model of retrieval-induced forgetting (RIF)

IOP Conf. Series: Journal of Physics: Conf. Series **1096** (2018) 012079 doi:10.1088/1742-6596/1096/1/012079

Psychological perspectives on implicit regularization: A model of retrieval-induced forgetting (RIF)

I Kulikovskikh[1], **S Prokhorov**[1]

[1]Samara National Research University, MoskovskoeShosse 34, Samara, Russia, 443086

e-mail: kulikovskikh.i@gmail.com, sp.prokhorov@gmail.com

Abstract. While learning theory suggests different forms of explicit regularization to guarantee small generalization error, deep learning models may introduce some sort of implicit regularization that tries to find a solution with small complexity, but neither does not include a penalty term nor does not directly modifies the optimization method. Recent research in deep learning pointed to the importance of proper understanding the underlying mechanisms of implicit regularization to elucidate the nature of generalization ability. This study is aimed at looking at implicit regularization from a psychological perspective with regard to 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.

1. Introduction

Deep learning allows highly efficient models that have improved the state-of-the-art in a broad range of applications [1,2]. 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 [3–5]. However, commonly used in deep learning gradient-based algorithms still require thorough understanding to overcome a variety of difficulties and limitations [6–8]. In particular, the vanishing gradient issue in first-order optimization methods seems impossible to tackle due to its inability to store and latch on the data structure for long time lags [9–13]. To compensate for the lack of an external memory, the recent studies suggest different approaches: a model that remembers old tasks by selectively slowing down learning on the weights important for those tasks [12]; a differentiable neural computer that can read from and write to an external memory matrix [10]; an easy-to-implement technique of adding gradient noise to very deep learning models [14,15]; a optimization method combined with various forms of well-chosen initialization (guessing weights) and momentum-based acceleration [16] and so on. Finally, there seems a problem of distinguishing between deep learning models that have different generalization performance. The traditional approaches like VC dimension, Rademacher complexity, uniform stability fail to explain why these models may generalize well in practice [17].

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 [16–21] 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 [21] 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. For this purpose, the present research models the phenomena of retrieval-induced forgetting (RIF) based on the extended logistic regression with respect to forgetting and guessing factors [22, 23]. Following from the definitions of floor and ceiling effects in statistics [23,24] and psychology [25,26], these factors allow a RIF model that helps to improve the convergence of gradient-based methods and, thus, to control the generalization ability.

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 [23, 24]. 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 [25,26]. 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 [27, 28]. 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 [27,29], 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) [27, 29], 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 [30, 31].

Thus, to address the problem of proper assessment in presence of floor and ceiling factors, Macready and Dayton [32] made two underlying assumptions:

(i) an observed failure in test results stems from forgetting;

(ii) an observed success in test results is attributed to guessing.

Following this, let us propose a RIF model based on forgetting and guessing factors.

3. A RIF model

Let $(x_i, y_i)_{i=1}^m$ be independent and identically distributed observations with responses $y_i \in \{0, 1\}$. Then, for any vector $\theta \in R^n$ of coefficients logistic regression models the class conditional probabilities $p(x_i, \theta) = P(y_i = 1|x_i, \theta)$ by

$$\ln\left(\frac{p(x_i,\theta)}{1-p(x_i,\theta)}\right) = \theta^T x_i.$$

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 $c^{g,f} = (c^g, c^f)$, where $c^{g,f} \in [0, 1]$. Using these denotations the logit function may be extended with regard to $c^{g,f}$ as

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

where

$$p(x_i, \theta, c^{g,f}) = c^g + \frac{1-(c^g+c^f)}{1+\exp(-\theta^T x_i)}. \tag{2}$$

Taking into account the extended definitions (1) and (2), let us introduce a RIF model $c^{g,f*}_{\{m,n\}}$ with respect to $\{m, n\} \in N$ in the form:

$$c^{g,f*}_{\{m,n\}} = \ln L(\theta, c^{g,f})_{\{m,n\}} \xrightarrow[c^{g,f}\in[0,1]]{} \min, \tag{3}$$

where the extended logistic loss function is given by

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

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 model.

4. Experiments

A brief description of five datasets used to validate the theoretical results is given in table 1. The information includes the values of and the class distribution. These datasets are freely available from UCI Machine Learning repository.

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 $m \in \{ak + b|k \in [0, K]\}$, where the number of steps $K = 4$. The rate of RIF 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:

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

and

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

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 ($0 \leq \varepsilon_{\{m,n\}} < 1$) with little improvement ($-0.1 \leq \varepsilon_{\{m,n\}} < 0$) in borderline cases.

Finally, BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm [33] was implemented to find a solution

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

BFGS's gradient updates involved computing $\nabla \ln L(\theta, c^{g,f})_{\{m,n\}}$ to approximate the Hessian matrix. The detailed analysis of the estimates' bias and the issues of convexity and convergence are beyond the scope of this paper, but still seems promising direction for further research.

Table 1. A brief description of the datasets

	m			n
dataset	$y_i \in \{0,1\}$	$y_i = 0$	$y_i = 1$	
Breast Cancer Wisconsin (breast-win)	683	444	239	9
Heart Statlog (heart)	270	150	120	13
Vertebral Column (vertebral)	309	100	209	6
Liver Disorder (liver)	345	145	200	6
Pima Indians Diabetes (pima)	768	500	268	8

Having discussed the design of computational experiments, let us define $\Delta_{\{m,n\}}$ and $\varepsilon_{\{m,n\}}$ based on cross-validation estimates of prediction error $\ln L(\theta, c^{g,f})_{\{m,n\}}$ (see table 2). 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 model is an attempt to model short-term memory with a guessing technique. To keep the positive trend in $\delta_{\{m,n\}}$, the proposed RIF model (3) needs to be extended to the case of long-term memory.

Figures 1-5 show the models of RIF $c^{g,f}_{\{m,n\}}$ for each dataset. The steps of both training and cross-validation traces are presented as circles the radius of which is increasing at the rate of RIF $\{a, b\}$ in $m \in \{ak + b | k \in [0, K]\}$.

5. Conclusion

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 a RIF model. The in-depth analysis showed that the inclusion of this model into gradient-based methods results in improved convergence of log-likelihood and depends on the relationship between forgetting

Table 2. Cross-validation estimates of prediction error

dataset	m	$c^{g,f}_{\{m,n\}}$	$\ln L(\theta, c^{g,f})_{\{m,n\}}$	$\delta_{\{m,n\}}$	$\varepsilon_{\{m,n\}}$
breast-win	20	(0.0735,0.1029)	0.03262	0.88577	-
($\varepsilon_m < 0.1$)	35	(0.0784,0.1175)	0.07126	0.82437	0.04687
	50	(0.0784,0.1175)	0.10551	0.74073	0.04623
	65	(0.0784,0.1175)	0.13224	0.68204	0.03919
	80	(0.0784,0.1175)	0.15008	0.64315	0.02773
heart	20	(0.0196,0.1029)	0.3136	0.72983	-
($\varepsilon_m < 1$)	35	(0.0196,0.1004)	0.47496	0.6602	0.24441
	50	(0.0196,0.1029)	0.60689	0.58662	0.2249
	65	(0.0637,0.1151)	0.70763	0.51334	0.19624
	80	(0.0196,0.0661)	0.77602	0.43998	0.15543
vertebral	20	(0.0588,0.1151)	0.45669	0.35931	-
($\varepsilon_m < -0.01$)	35	(0.0049,0)	0.62598	0.19335	0.87553
	50	(0.0049,0)	0.63339	0.15636	0.04741
	65	(0.0049,0)	0.63188	0.1402	**-0.01079**
	80	(0.0049,0)	0.62239	0.12259	**-0.07739**
liver	20	(0,0.0073)	0.81139	0.07104	-
($\varepsilon_m < -1$)	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	20	(0.0955,0.1176)	2.25079	0.19788	-
($\varepsilon_m < -1$)	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**

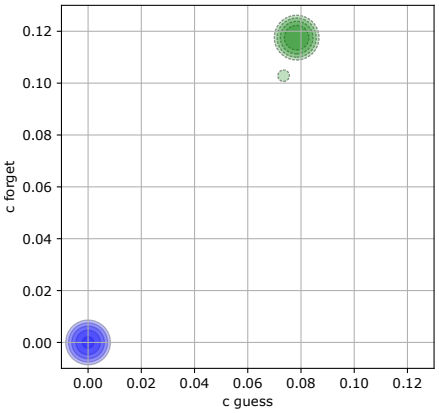

Figure 1. Breast-win

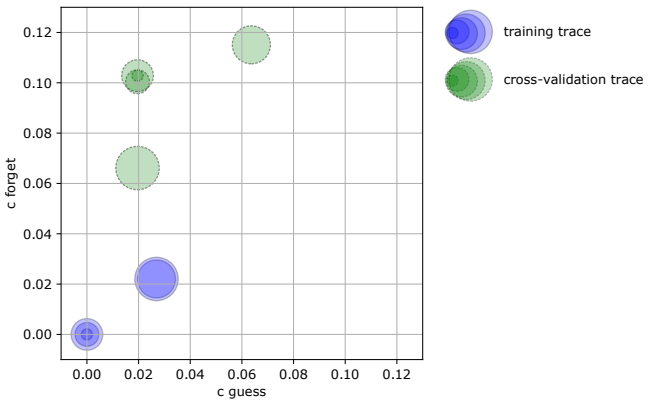

Figure 2. Heart

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.

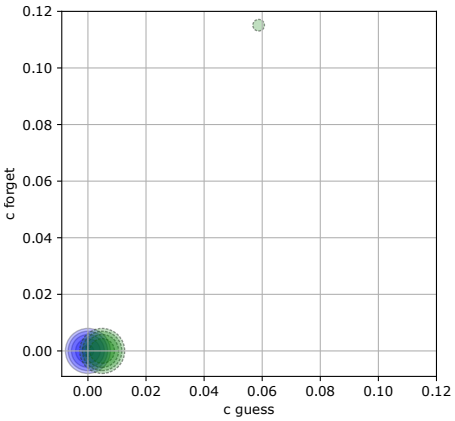

Figure 3. Vertebral

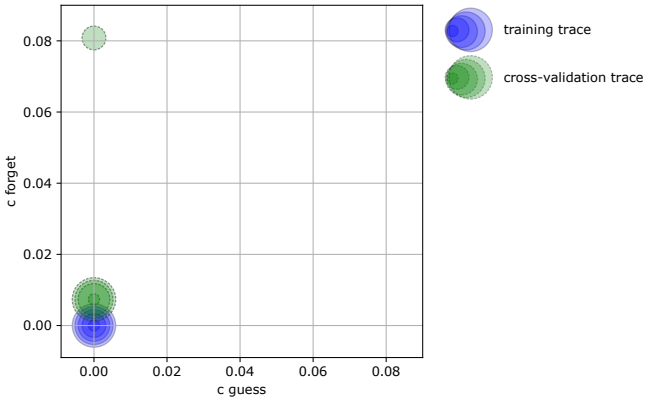

Figure 4. Liver

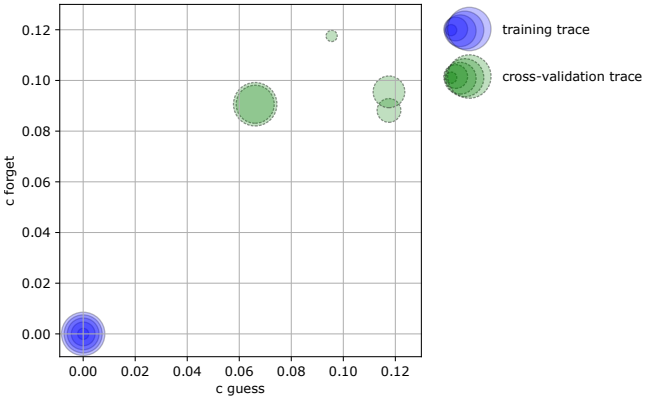

Figure 5. Pima

6. References

[1] Spitsyn VG, Bolotova YuA, Phan NH and Bui TTT 2016 Using a Haar wavelet transform, principal component analysis and neural networks for OCR in the presence of impulse noise *Computer Optics* **40(2)** 249-257 DOI: 10.18287/2412-6179-2016-40-2-249-257

[2] Savchenko AV 2017 Maximum-likelihood dissimilarities in image recognition with deep neural networks *Computer Optics* **41(3)** 422-430 DOI: 10.18287/2412-6179-2017-41-3-422-430

[3] LeCun Y, Bengio Y and Hinton G 2015 *Nature* **521** 436-44

[4] Ronen R 2017 Why & When Deep Learning Works: Looking Inside Deep Learnings *The Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI)*

[5] Shwartz-Ziv R and Tishby N 2017 Opening the Black Box of Deep Neural Networks via Information *Machine Learning* arXiv:1703.00810

[6] Bengio Y, Simard P and Franconi P 1994 *IEEE Trans. Neural Networks* **5(2)** 157-166

[7] Hochreiter S 1996 *Int. J. Uncertainty Fuzziness Knowl.-Based Syst.* **6(2)** 107-116

[8] Shalev-Shwartz S, Ohad Shamir O and Shammah S 2017 Failures of Gradient-Based Deep Learning *Machine Learning* arXiv:1703.07950

[9] Arpit D, Jastrzebski S, Ballas N, Krueger D, Bengio B, Kanwal M S, Maharaj T, Fischer A, Courville A, Bengio Y and Lacoste-Julien S 2017 A closer look at memorization in deep networksar *Proceedings of the 34th International Conference on Machine Learning (ICML)* Xiv:1706.05394

[10] Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, Grabska-Barwinska A, Colmenarejo S G,Grefenstette E, Ramalho T, Agapiou J, Badia A P, Hermann K M, Zwols Y, Ostrovski G, Cain A, King H, Summerfield C, Blunsom P, Kavukcuoglu K and Hassabis D 2016 *Nature* **538** 471-476

[11] Hochreiter S and Schmidhuber J 1997 *Neural Comput.* **9** 1735-1780
[12] Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu A A, Milan K, Quan J, Ramalho T, Grabska-Barwinska A, Hassabis D, Clopath C, Kumaran D and Hadsell R 2017 Overcoming catastrophic forgetting in neural networks *Proceedings of the 34th International Conference on Machine Learning (ICML)* arXiv:1612.00796
[13] Santoro A, Bartunov S, Botvinick M, Wierstra D and Lillicrap T 2016 One-shot Learning with Memory-Augmented *Neural Networks Machine Learning* arXiv:1605.06065
[14] Bishop C M 1995 *Neural Comput.* **7(1)** 108-116
[15] Neelakantan A, Vilnis L, Le Q V, Sutskever I, Kaiser L, Kurach K and Martens J 2015 Adding Gradient Noise Improves Learning for Very Deep Networks *Machine Learning* arXiv:1511.06807
[16] Fan Q, Wu W and Zurada J M 2016 Convergence of batch gradient learning with smoothing regularization and adaptive momentum for neural networks *Springerplus* **5** 295 DOI: 10.1186/s40064-016-1931-0
[17] Zhang C, Bengio S, Hardt M, Recht B and Vinyals O 2017 Understanding deep learning requires rethinking generalization *Machine Learning* arXiv:1611.03530
[18] Gunasekar S, Woodworth B, Bhojanapalli S, Neyshabur B and Srebro N 2017 Implicit Regularization in Matrix Factorization *Machine Learning* arXiv:1705.09280
[19] Neyshabur B 2017 Implicit Regularization in Deep Learning *Machine Learning* 1-110 arXiv:1709.01953
[20] Neyshabur B, Tomioka R and Srebro N 2014 In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning *Machine Learning* arXiv:1412.6614
[21] Neyshabur B, Tomioka R, Salakhutdinov R and Srebro N 2017 Geometry of Optimization and Implicit Regularization in Deep Learning *Machine Learning* arXiv:1705.03071
[22] Kulikovskikh I M 2017 Cognitive validation map for early occupancy detection in environmental sensing *Eng. Appl. Artif. Intell.* **65** 330-335
[23] Kulikovskikh I M and Prokhorov S A 2017 *Procedia Eng.* **201** 779-788
[24] Donnelly S and Verkuilen J 2017 *J. Mem. Lang.* **94** 28-42
[25] Everitt B S 2010 *The Cambridge Dictionary of Statistics* (Cambridge: Cambridge University Press)
[26] Groth-Marnat G and Wright A J 2016 *Handbook of Psychological Assessment* (Wiley)
[27] Bjork E L, Soderstrom N C and Little J L 2015 *Am. J. Psychol.* **128(2)** 229-239
[28] Elliott G, Isaac C L and Muhlert N 2014 *Cortex* **54** 16-32
[29] Little J L, Bjork E L 2015 *Mem. Cogn.* **43** 14-26
[30] Chan J C K 2009 *J. Mem. Lang.* **61(2)** 153-170
[31] Kubinger K D, Holocher-Ertl S, Reif M , Hohensinn C, Frebort M 2010 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 *Int. J. Sel. Assess.* **18(1)** 111-115
[32] Macready G B, Dayton C M 1977 *J. Educ. Stat.* **2** 99-120
[33] Fletcher R 1987 *Practical methods of optimization* (New York: John Wiley)

Acknowledgments
This work was supported by the Russian Federation President grant MK-6218.2018.9, the Ministry of Education and Science of the Russian Federation grant 074-U01 and supported in part by Russian Foundation for Basic Research (project No. 18-37-00219).