NEURONS LEARN SLOWER THAN THEY THINK

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies revealed complex convergence dynamics in gradient-based methods, which has been little understood so far. Changing the step size to balance between high convergence rate and small generalization error may not be sufficient: maximizing the test accuracy usually requires a larger learning rate than minimizing the training loss. To explore the dynamic bounds of convergence rate, this study introduces *differential capability* into an optimization process, which measures whether the test accuracy increases as fast as a model approaches the decision boundary in a classification problem. The convergence analysis showed that: 1) a higher convergence rate leads to slower capability growth; 2) a lower convergence rate results in faster capability growth and decay; 3) regulating a convergence rate in either direction reduces differential capability.

1 INTRODUCTION

Training a model is an optimization problem that involves minimizing the model errors on a training dataset. When training a network with gradient-based methods, accelerating convergence to the solution is of a top priority (Dieuleveut et al., 2017; Arora et al., 2019), but not the only performance variable to optimize. Minimizing the difference between the model errors on a training subset and a testing subset, which is called the generalization error, plays a fundamental role (Hardt et al., 2016; Zhang et al., 2017; Lin, 2019).

While adapting a step size in iterative optimization schemes increases the convergence rate, it does not deliver the best generalization error (Luo et al., 2019; Xie et al., 2020; Heo et al., 2021; Zhou et al., 2020). Adaptive optimization methods are often outperformed by non-adaptive stochastic gradient descent (SGD) for overparameterized models, where the number of trainable parameters is much higher than the number of samples they are trained on. Recent studies revealed that overparametrization itself leads to faster convergence (Arora et al., 2018; Li & Liang, 2018; Allen-Zhu et al., 2019; Oymak & Soltanolkotabi, 2019; Liu & Belkin, 2020; Oymak & Soltanolkotabi, 2020; Chen et al., 2021). Besides, a step size on testing is usually larger than a step size on training (Bortoli et al., 2020; Li & Arora, 2020; Cohen et al., 2021). These findings point to the fact that convergence demonstrates more complex dynamics which has not been well understood so far.

The study raises a research question on whether we can deepen our understanding by inspecting human and machine reasoning processes in a testing environment. Originating from the item response theory (IRT) (Lord, 1980; de Ayala, 2009), *differential capability* shows how fast increases the probability of a correct response to an item of a given difficulty in comparison with a learner's ability. With a new interpetation in a machine learning context, the differential capability may identify how fast increases the test accuracy compared to a model's ability to reach a decision boundary in a classification problem. It seems to be an appropriate measure to answer our research question. The work related to this problem is briefly discussed in Appendix A.

2 DIFFERENTIAL CAPABILITY

2.1 PROBLEM SETUP

For a dataset $\{\mathrm{x}_i, y_i\}_{i=1}^m$ with $\mathrm{x}_i \in \mathrm{R}^n$, $y_i \in \{-1, 1\}$, let us minimize an empirical loss function with a weight vector $\theta \in \mathrm{R}^n$:

$$\mathcal{L}(\theta) = \sum_i \ell(y_i \theta^\top \mathrm{x}_i), \tag{1}$$

where ℓ measures the discrepancy between the output y and the model prediction. The gradient descent (GD) finds the weight vector with a fixed step size η:

$$\theta(t+1) = \theta(t) - \eta \nabla_{\theta} \mathcal{L}(\theta). \tag{2}$$

For a large family of monotone losses with polynomial and exponential tails (Nacson et al., 2019), the derivative of $\ell(t)$ can be presented as $\ell'(t) = -e^{-f(t)}$, where $f(t)$ satisfies $\forall k \in \mathrm{N}$: $\left|\frac{f^{k+1}(t)}{f'(t)}\right| = \mathcal{O}(t^{-k})$. The continuous form of equation 2 ($\eta \to 0$) is equal to $\theta'(t) = \sum_i e^{-f(y_i \mathrm{x}_i^{\top} \theta(t))} y_i \mathrm{x}_i$, where the weight vector can be presented asymptotically as $\theta(t) = g(t)\hat{\theta} + h(t)$, $h(t) = o(g(t))$, where $g(t)$ defines a convergence rate, $\hat{\theta} = \arg\min_{\theta \in \mathrm{R}^n} \|\theta\|^2$, so that $y_i \theta^{\top} \mathrm{x}_i \geq 1$ (Soudry et al., 2018; Nacson et al., 2019) . Using $\ell'(t)$, we can write:

$$g'(t)\hat{\theta} = \sum_i e^{-f(g(t) y_i \mathrm{x}_i^{\top} \hat{\theta} + h(t) y_i \mathrm{x}_i^{\top}} y_i \mathrm{x}_i \approx e^{-f(g(t)} \sum_i e^{-f'(g(t))h(t) y_i \mathrm{x}_i^{\top}} y_i \mathrm{x}_i.$$

For the last equation, we can require $g'(t) = e^{-f(g(t)}$. Approximating it with $g'(t) \approx e^{-f(g(t)) - \ln f'(g(t))}$ gives us a closed from solution $g(t) = f^{-1}(\ln t + C)$.

The present study enriches the provided reasoning with differential capability, which measures whether the test accuracy increases as fast as a model's ability to reach a decision boundary. Introducing differential capability into an optimization process, this research explores the dynamic bounds of a convergence rate and reveals how an increase/decrease in a convergence rate affects the proposed measure.

Related work often attributed the success in balancing convergence and generalization to the complexity and capacity of neural networks. To observe the impact of differential capability on convergence dynamics, which is not affected by network architecture, the present study focuses on the simplest learner model - a single neuron, the capacity of which stimulates the renewed interest (Frei et al., 2020; Gidon et al., 2020; Jones & Kording, 2020; Yehudai & Shamir, 2020).

2.2 Loss function with differential capability

Let us build a loss function on the well-studied two-parameter logistic item response theory (2PL IRT) model Lord (1980); de Ayala (2009): $P(y_{ij} = 1|\omega_j, r_i, d_i) = \frac{1}{1+\exp(-r_i(\omega_j - d_i))}$, where $P(y_{ij} = 1|\omega_j, r_i, d_i)$ is a probability of correctly responding $y_{ij} = 1$ to an item i with a difficulty d_i by a learner j with the ability ω_j; r_i is a discrimination parameter that measures the differential capability of an item i. A high value of r_i means that the probability of a correct response to an item with a given difficulty increases as quickly as a learner's ability. When $r_i = 1$ and $d_i = 0$, the 2PL IRT model reduces to the sigmoid function.

In comparison with the 2PL IRT model, the new loss function, equipped with differential capability, changes the shape of the sigmoid so that it becomes non-monotonic and exhibits more complex behavior. First, differential capability grows with a rate r_i. When a learner acquires the ability ω_j to respond correctly to an item with difficulty d_i, the differential capability decays with a rate c_i. The probability of answering correctly to an item is equal to P_{d_i}. Using the Gompertz logistic law of population dynamics (Gray & Gray, 2017), a learner's response to an item i can be defined as:

$$P(y_{ij} = 1|\omega_j, r_i, a_i, b_i, d_i) = a_i e^{b_i e^{-r_i(\omega_j - d_i)}}, \tag{3}$$

where $a_i = e^{\varepsilon_i}$, $\varepsilon_i = \frac{c_i}{r_i}$, $b_i = \ln P_{d_i} - \varepsilon_i$. The parameter ε_i reflects the balance between a growing rate r_i and a decaying rate c_i for an item i. Fig. 1 depicts different configurations of the DC loss function, where DC stands for differential capability.

2.3 Convergence analysis

Interpreting differential capability in a machine learning context, where a high value of this measure points out that the test accuracy increases as fast as a model ability to reach a decision boundary, the equation 3 can be rewritten as $\ell^{DC}(t) = ae^{be^{-r(t-d)}}$, for which $\ell'^{DC}(t) = -abre^{-f(t)}$, where

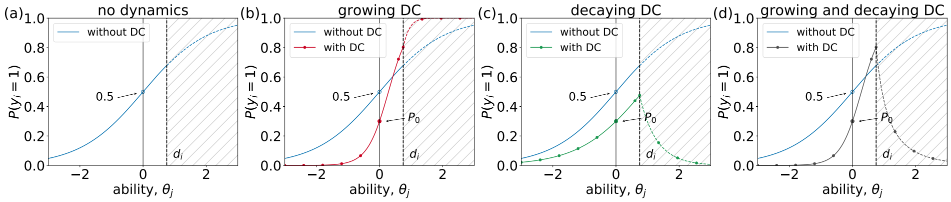

Figure 1: Different configurations of the DC loss function: (a) no DC ($r_i = 1, c_i = 0$), (b) growing DC ($r_i > 0, c_i = 0$), (c) decaying DC ($r_i = 1, c_i > 0$), and (d) both growing and decaying DC ($r_i > 0, c_i > 0$).

$f(t) = r(t-d) - be^{-r(t-d)}$. According to the reasonong presented in Section 2.1, estimating the inverse function of $f(t)$ gives the convergence rate:

$$g^{DC}(z) = d + \big(\mathrm{W}_0(be^{-z}) + z\big)/r, \quad z > 0, \tag{4}$$

where $\mathrm{W}_0(z)$ is the principal branch of the Lambert function. Nacson et al. (2019); Soudry et al. (2018) showed that for any strict monotone loss $\ell(t)$, given in Section 2.1, under certain conditions, $g(t) = \ln t$. With a variable substitute $z = \ln t$, the convergence rate $g(z) = z$ is further refered to as the default rate.

Let us first analyze the parameters b, d, and r, which affect the convergence rate equation 4. We can see that the difficulty denoted by $d > 0$ increases the absolute value of $g^{DC}(z)$. The parameter b depends on P_d and the ratio c/r: $b = \ln P_d - c/r$ (see equation 3). As $0 < P_d < 1$, $\ln(P_d) < 0$. The smaller P_d is, the faster $|\ln P_d|$ increases. The ratio $c/r > 0$ grows up if $r \to 0$ (an infinitesimal value) or/and $c > r$. The parameter $b < 0$, but smaller P_d, r and larger c increase its absolute value.

Let us now show that differential capability dynamically changes the bounds of convergence rate.

Theorem. *For any $z > 0$, $b < 0$, moderate $r > 0$, and $d = 0$, the bounds of the convergence rate $g^{DC}(z)$ given by equation 4 are below and above the default convergence rate $g(z) = z$.*

The proof of the theorem is deferred to Appendix B.

Corollary. *The bounds of $g^{DC}(z)$ move to the left when r is larger and to the right when d is larger and r is smaller.*

Proof. The validity of the corollary follows from equation 4. □

From the convergence analysis, we can conclude that: 1) a higher convergence rate leads to a lower growth rate r, which results in smaller differential capability; 2) a lower convergence rate leads to a higher growth rate r, which is compensated by a higher decay rate c, and, thus, results in smaller differential capability again. This means that regulating a convergence rate in either direction does not increase differential capability.

3 Experimental results

As the proposed measure dynamically changes the bounds of the convergence rate $g(t)$, it also brings more flexibility in regulating the trade-off between a convergence rate and an error rate. Fig. 2 illustrates how differential capability affects the inner processes inside a neuron for the loss configurations given in Fig. 1. It replaces the superposition in equation 1 with more complex dynamics (see Fig. 2 (a), (b) in comparison with Fig. 2 (c)-(h)), where the balance between a growth rate r and a decay rate c regulates the convergence/error rate trade-off.

Let us adopt this illustration to design the experiments on a set of synthetic datasets ($m = 1000$, $n = 2$), which were randomly split into training (80%) and testing (20%) subsets (see Fig. 3, in the upper left corner). A neuron adjusted its weights with gradient descent in a stochastic setting (SGD) with the default parameters, batch size $|B(t)| = 75$, and $n_{\text{epoch}} = 1500$. The number of runs was equal to 10. The hyperparameters were chosen within the following regions: $d \in [0, 5]$, $P_d \in [0.1, 0.9]$, $r \in [0.1, 12]$, and $c \in [0, 12]$ with a 2.5% random pick from the full grid space. In Fig. 3, the label “no DC” reflects the default configuration with the sigmoidal loss function (see 2 (a), (b)), “DC $r \downarrow$” denotes the configuration with growing DC 2 (c), (d)), and “DC $r \uparrow$, $c \uparrow$”

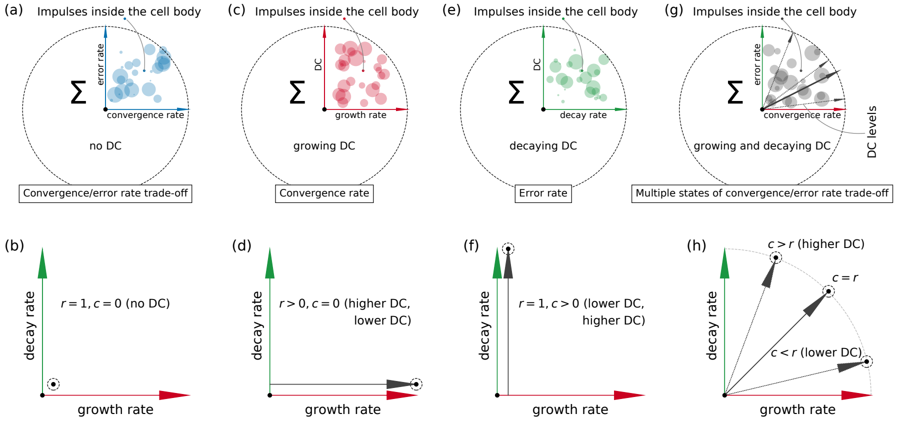

Figure 2: The impact of differential capability (DC) on inner processes inside a neuron

stands for the configuration with growing and decaying DC 2 (g), (h)). Here, the configuration with only decaying DC was left with little attention. According to the provided convergence analysis, increasing c affects the dynamics bounds of convergence rate to a smaller extent (see Section 2.3).

The theoretical analysis revealed that a slower growth rate $r\downarrow$ reduces differential capability while increasing the convergence rate. Fig. 3 illustrates this result. The growth rate r increases slower than it needs to reach the highest test accuracy. But, its value $r > 1$, which means it naturally enlarges the step size of the optimizer (see Section 2.3, $\ell'^{DC}(t)$). As a consequence, we can observe the higher test accuracy (see the red curves in contrast to the blue ones on the plots). This is exactly the phenomenon this study is intended to demystify.

A faster growth rate $r\uparrow$ and decay rate $c\uparrow$ reduce differential capability as well while decreasing the convergence rate. A non-zero value of c balances against even higher value of r, which, on the one hand, slows down the convergence, on the other hand, increases the test accuracy to a greater extent (see the black curves in contrast to the red and blue ones on the plots). From the above empirical analysis, we can conclude that neither increase nor decrease in a convergence rate improves differential capability as the model does not achieve the highest test accuracy.

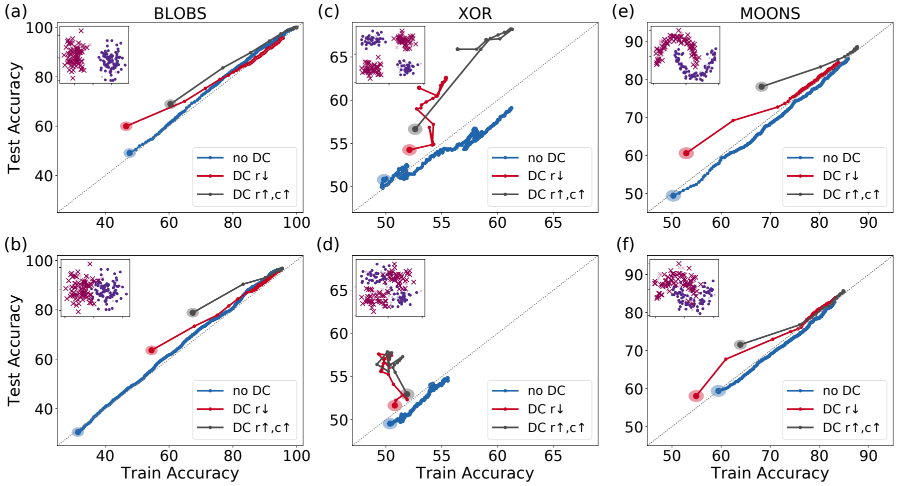

Figure 3: The impact of differential capability (DC) on convergence/error rate trade-off

4 Conclusion

This study explored the dynamic bounds of a convergence rate with differential capability, which measures how fast increases the test accuracy compared to a model's ability to reach a decision boundary in a classification problem. The provided analysis enriched the understanding of convergence dynamics and revealed that both increase and decrease in a convergence rate reduce differential capability.

References

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *ICML*, 2019.

S. Arora, N. Cohen, and E. Nazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *ICML*, 2018.

S. Arora, N. Cohen, N. Golowich, and W. Hu. A convergence analysis of gradient descent for deep linear neural networks. In *ICLR*, 2019.

V. De Bortoli, A. Durmus, X. Fontaine, and U. Simsekli. Quantitative propagation of chaos for sgd in wide neural networks. In *NeurIPS*, 2020.

Y. Chen, T. S. Filho, R. B. C. Prudencio, T. Diethe, and P. Flach. β^3-irt: A new item response model and its applications. In *AISTATS*, 2019.

Z. Chen, Y. Cao, D. Zou, and Q. Gu. How much over-parameterization is sufficient to learn deep relu networks? In *ICLR*, 2021.

J. Cohen, S. Kaur, Y. Li, Z. Kolter, and A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *ICLR*, 2021.

R. J. de Ayala (ed.). *The Theory and Practice of Item Response Theory (Methodology in the Social Sciences)*. The Guilford Press, New York, 2009.

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18:1–51, 2017.

S. Frei, Y. Cao, and Q. Gu. Agnostic learning of a single neuron with gradient descent. In *NeurIPS*, 2020.

A. Gidon, T. Adam Zolnik, P. Fidzinski, F. Bolduan, A. Papoutsi, P. Poirazi, M. Holtkamp, I. Vida, and M. E. Larkum. Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science*, 367(6473):83–87, 2020. doi: 10.1126/science.aax6239.

M. J. Gierl, O. Bulut, Q. Guo, and X. Zhang. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87 (6):1082–1116, 2017. doi: 10.3102/0034654317726529.

W. G. Gray and G. A. Gray (eds.). *Introduction to Environmental Modeling*. Cambridge University Press, Cambridge, UK, 2017.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, 2016.

B. Heo, S. Chun, S. Joon Oh, D. Han, S. Yun, G. Kim, Y. Uh, and J.-W. Ha. Adamp: Slowing down the slowdown for momentum optimizers on scale-invariant weights. In *ICLR*, 2021.

I. S. Jones and K. Kording. Can single neurons solve mnist? the computational power of biological dendritic trees. *arXiv preprint*, arXiv:2009.01269v1, 2020.

I. Kulikovskikh. Cognitive validation map for early occupancy detection in environmental sensing. *Engineering Applications of Artificial Intelligence*, 65:330–335, 2017. doi: https://doi.org/10.1016/j.engappai.2017.08.008.

I. Kulikovskikh, T. Lipic, and T. Šmuc. From knowledge transmission to knowledge construction: A step towards human-like active learning. *Entropy*, 22(8), 2020. doi: 10.3390/e22080906.

J. P. Lalor, H. Wu, T. Munkhdalai, and H. Yu. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4711–4716, 2018. doi: 10.18653/v1/D18-1500.

Y. Li and Y. Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *NeurIPS*, 2018.

Z. Li and S. Arora. An exponential learning rate schedule for deep learning. In *ICLR*, 2020.

D. Liang, F. Ma, and W. Li. New gradient-weighted adaptive gradient methods with dynamic constraints. *IEEE Access*, 8:110929–110942, 2020. doi: 10.1109/ACCESS.2020.3002590.

S. Lin. Generalization and expressivity for deep nets. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5):1392–1406, 2019. doi: 10.1109/TNNLS.2018.2868980.

C. Liu and M. Belkin. Accelerating sgd with momentum for over-parameterized learning. In *ICLR*, 2020.

F. M. Lord (ed.). *Applications of Item Response Theory To Practical Testing Problems*. Routledge, New York, 1980.

L. Luo, Y. Xiong, Y. Liu, and X. Sun. Adaptive gradient methods with dynamic bound of learning rate. In *ICLR*, 2019.

F. Martínez-Plumed, R. B. C. Prudêncio, A. Martínez-Usó, and J. Hernández-Orallo. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42, 2019. doi: https://doi.org/10.1016/j.artint.2018.09.004.

M. S. Nacson, J. Lee, S. Gunasekar, N. Srebro, and D. Soudry. Convergence of gradient descent on separable data. In *AISTATS*, 2019.

S. Oymak and M. Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *ICML*, 2019.

S. Oymak and M. Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020. doi: 10.1109/JSAIT.2020.2991332.

D. Soudry, E. Hoffer, M.S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *JMLR*, 19:1–57, 2018.

S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *NeurIPS*, 2019.

J. Wu, D. Zou, V. Braverman, and Q. Gu. Direction matters: On the implicit regularization effect of stochastic gradient descent with moderate learning rate. In *ICLR*, 2021.

Y. Xie, X. Wu, and R. Ward. Linear convergence of adaptive stochastic gradient descent. In *AISTATS*, 2020.

G. Yehudai and O. Shamir. Learning a single neuron with gradient methods. In *COLT*, 2020.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Y. Zhou, B. Karimi, J. Yu, Z. Xu, and P. Li. Towards better generalization of adaptive gradient methods. In *NeurIPS*, 2020.

A Related work

The analysis of SGD based optimization for overparameterized models has recently become an active area of research interest (Arora et al., 2018; Li & Arora, 2020; Arora et al., 2019; Allen-Zhu et al., 2019; Zhou et al., 2020; Wu et al., 2021). Recent studies indicated that large learning rates can preserve good generalization and accelerate SGD convergence with no additional gradient scaling. While analyzing the effect of overparametrization, Wu et al. (2021) pointed to the difference in directional biases for SGD and GD with a moderate and annealing rates. Vaswani et al. (2019) explored line-search techniques and provided heuristics to automatically set larger learning rates. Li

& Arora (2020) analyzed an exponential rate schedule. They showed that using SGD with momentum (Liu & Belkin, 2020) and an exponentially increasing rate, coupled with batch normalization, maintains a good balance between convergence and generalization across all standard architectures.

In line with more successful SGD adoption for overparameterized models, remarkable progress has been achieved in optimization methods with adaptive learning rates. SGDP and AdamP use effective rates without changing the update directions (Heo et al., 2021), which allows preserving the original convergence properties of GD optimizers. RAdam adopts the learning rate warm-up heuristic to rectify the variance of adaptive rates (Liu & Belkin, 2020) and, by that, stabilize training, accelerate convergence, and improve generalization. To balance generalization and convergence on unstable and extreme learning rates, Luo et al. (2019) put forward AdaBound and AMSBound which adopt dynamic bounds on rates to eliminate the generalization gap between adaptive methods and SGD and maintain a higher learning rate early in the training. These methods were further developed with regard to a dynamic decay rate in (Liang et al., 2020).

Adopting the item response theory to address interpretability and explainability issues in deep learning has been reported to be successful. Kulikovskikh (2017) extended the model of logistic regression with 4PL IRT model to reduce the disruptive influence of floor and ceiling effects on the convergence of log-likelihood. Lalor et al. (2018) inverstigated the relationship between items difficulty and model performance in deep networks. Martínez-Plumed et al. (2019) interpretated the IRT model parameters in terms of a classification problem. Chen et al. (2019) proposed a new IRT model, which allows simulating continuous responses and enriches the family of Item Characteristic Curves. The authors applied the model to evaluating the quality of different machine learning classifiers with items difficulty and discrimination. Kulikovskikh et al. (2020) suggested a new query strategy for an active learning environment to increase the transparency of deep network architectures.

B Proof of the theorem

Proof. For $z > 0$, the equation $we^w = z$ has one positive solution $w = \mathrm{W}_0(z)$, which increases with z. If $z = e$, then $w = 1$. Thus, $w > 1$ if $z > e$. By taking logarithms of both sides, we get:

$$\begin{aligned} \ln w + w &= \ln z; \\ w &= \ln z - \ln w < \ln z. \end{aligned} \tag{5}$$

When $z > e$,

$$\begin{aligned} 1 &< w < \ln x \\ 0 &< \ln w < \ln \ln z. \end{aligned} \tag{6}$$

Substituting equation 6 into equation 5 yields:

$$\ln z - \ln \ln z < w < \ln z, \tag{7}$$

where the left side is positive for $z > 1$. Since $w = \mathrm{W}_0(z)$, we can write:

$$\ln z - \ln \ln z < \mathrm{W}_0(z) < \ln z, \tag{8}$$

Let us now modify the argument of $\mathrm{W}_0(z)$ with regard to $g^{DC}(z)$:

$$\begin{aligned} \frac{b}{z} + z < \mathrm{W}_0(be^{-z}) + z &< \frac{b}{z} - \frac{b}{\ln z} + z; \\ \frac{b}{z} + z < \mathrm{W}_0(be^{-z}) + z &< b\frac{\ln z - z}{z \ln z} + z, \end{aligned}$$

where $\ln z - z < 0$ as $\ln z < z$ for all $z > 0$.

By definition, $b < 0$. Thus, for $z > e$:

$$\begin{aligned} b\frac{\ln z - z}{z \ln z} + z &> z; \\ \frac{b}{z} + z &< z \end{aligned}$$

As we see, the boundaries of $\mathrm{W}_0(be^{-z}) + z$ are below and above the default convergence rate z. □