Bachelor's Thesis

Survey on Regularization Methods in Continual Learning

Department of Statistics
Ludwig-Maximilians-Universität München

Jörg Schantz

Munich, March 20th, 2025

Submitted in partial fulfillment of the requirements for the degree of B. Sc.
Supervised by Dr. Julian Rodemann

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. [3]

Contents

1 Introduction

Bli bla bulb

2 Neural Networks

Although continual learning is general modeling concept, applicable in statistical inference as well as pattern driven prediction algorithms, it is mostly used a in machine learning context. More specifically in artificial neural networks (ANN). They are algorithms based on the functionality of a human brain and often designed for scenarios where data is seen in real-time e.g. stock market predictions or power control systems.
The simplest form of an ANN is a single linear classifier, called one-neuron perceptron, that devides a vector x into two classes using a so-called activation function $h(\cdot)$ [6]. The neuron's input is given by

$$\sum_{i=1}^{n} w_i x_i + c = w^\top x + c \tag{1}$$

where n is the number of observations, w a weight vector assigned to x and c the decision threshold. The two class regions are separated by the hyperplane [6]

$$w^\top x + c = 0 \tag{2}$$

. Using multiple neurons with the same activation function creates a one-layer perceptron and enables classification for more than two classes with the input

$$\sum_{k=1}^{m} \sum_{i=1}^{n} w_{k,i} x_i + c = (w_1^\top x + c, ..., w_m^\top x + c)^\top = W^\top x + c \tag{3}$$

where W is the $n \times m$ weight matrix and m the number of classes. Given h the logistic function a one-layer perceptron is equal to a multinomial logit model [10]. Composing l layers of neurons, Feed Forward NN (FFNN), allows for a more and more abstract representation of the data and finer class boundaries. The unknown weight matrices $W_1, ..., W_l$ and the decision threshold c are the solution to the minimization problem

$$\hat{\theta} = \arg\min \sum_{i=1}^{n} L(f(x_i, \theta), y_i) \tag{4}$$

where θ are the unknown parameters, and $L(\cdot)$ a loss function which measures the difference between the predicted values $f(x_i, \theta)$ and true values y_i.

3 Framework

Throughout literature continual learning in a statistical sense means modeling a joint probability distribution $\mathbb{P}^{(T)}$, which is allowed to expand indefinitely [25]. T samples $D_t, t \in 1, ..., T$ from different distributions $\mathbb{P}_t$ are processed sequentially. A single sample has the form $D_t = (x_i^{(t)}, y_i^{(t)})$ with $x_i^{(t)}$ being the i-th covariate and $y_i^{(t)}$ the dependent variable. The D_t are assumed to be conditionally independent but not necessarily identically distributed [25]. Each tuple $(D_t, \mathbb{P}_t)$ may correspond to a distinct regression or classification task that is to be learned. The goal is to train a single model which is able to perform well on all tasks, although it is trained sequentially and cannot necessarily revisit prior tasks.

3.1 Scenarios

In regards to the distribution $\mathbb{P}$ of $Y^{(t)} = \{Y_1, ..., Y_t\}$ over which the model is evaluated after seeing the t-th samples, [4] and [25] differentiate between eight CL scenarios:
Task-incremental learning (TIL), *Class-incremental learning* (CIL), *Task-Free continual learning* (TFCL) and *Online contiunal learning* (OCL) algorithms all aim to learn a distinct set of tasks, while providing a task identity, if not stated otherwise [4, 25].

$$\emptyset = Y_t \cap Y_{t+1} \Rightarrow \mathbb{P}(Y^{(t+1)}) = \prod_{i=1}^{t+1} \mathbb{P}(Y_i) \tag{5}$$

TIL allows task individual output layers or the training of separate models for each task. The challenge then is less about forgetting (subsection 3.2) but finding a healthy balance between predicting accuracy and model complexity [24].
CIL restricts this approach by only training one model, which is introduced stepwise to different classification tasks. CIL only provides task identity during training [24]. For example with samples t an agent learns to classify hats or gloves and with sample $t + 1$ shirts or pants. When testing, it is then also required to classify hats or shirts.
TFCL does not provide any task identity to the model and only focuses on labels [1].
OCL limits its sample sizes to one and focuses on real-time training [4, 25].
Domain-incremental learning (DIL) algorithms seek to learn multiple tasks that share the same label space [4]. For example first learning to drive during sunny weather and later

on while it is rainy.

$$Y_t = Y_{t+1} \nRightarrow \mathbb{P}(Y_t) = \mathbb{P}(Y_{t+1}) \tag{6}$$

One could view this as a version of task-incremental learning, where task identity is secondary as all tasks have the same data labels. Thus design based strategies to inhibit catastrophic forgetting are not possible [24].

Instance-incremental learning (IIL) algorithms learn one common task for all training samples [4, 25].

$$Y_t = Y_{t+1}, \mathbb{P}(Y_t) = \mathbb{P}(Y_{t+1}) \Rightarrow \mathbb{P}(Y^{(t+1)}) = \mathbb{P}(Y_1) \tag{7}$$

This is a special case of DIL where a model learns the distribution of one "domain" while only ever accessing snippets of the total available data. For example each sample contains new real-world photographs of cats to classify. Assuming OCL only learns one task, OCL is a special case of IIL where every data point is seen in sequence.

Blurred Boundary continual learning (BBCL), in contrast to all others so far, allows partially overlapping label spaces [4, 25].

Continual Pre-training (CPT) aims to improve knowledge transfer with sequentially arriving pre-training data [4, 25].

3.2 Stability-Plasticity Trade-off

The challenge of continual learning is to strike a balance between stability and plasticity. Models should retain knowledge of past tasks, stability, while being flexible enough to incorporate information from new data, plasticity. The sequential training nature of CL changes the weights acquired form learning task A to accommodate for a new task B. This abrupt loss of information is called catastrophic forgetting [12, 18, 19, 21]. A naive approach to solving this dilemma would be storing and replaying data to the network with each training step. This is impractical because the amount of data needed to be stored is proportional to the number of tasks learned.

Evron et al. define forgetting as

$$F(k) = 1/k \sum_{t=1}^{k} \|X_t w_k - y_t\|^2 \tag{8}$$

. They have analyzed catastrophic forgetting in linear regression under the assumptions that values of X are bounded by 1, tasks are jointly realizable with a bounded (by 1) norm and there are more parameters than observations in each sample. Realizability assumes the existence of true model weights s.t. $y = Xw$ [23]. This enables them to focus only on minimizing the distance between new and old model weights. In their work they find an

upper bound for forgetting

$$\sup F(k) = \sup 1/k \sum_{k}^{t=1} \|(I - Q_t)Q_k ... Q_1\|^2 \tag{9}$$

where Q_i are the projections onto the solution spaces of w_i i.e. $Q_i := I - X_i^\top (X_i X_i^\top)^{-1} X_i$. So far many methods of minimizing catastrophic forgetting have been developed. Their core ideas can be summarized to *Replay* methods [2, 5, 22], *Optimization* methods [13, 16, 20], *Architecual* methods [8, 11, 17] and *Regularization* methods, which will be discussed in section 5.

4 Metrics

Intro.
In the following each sample $D_t = (X_t, Y_t)$ is divided into a training split $D_t^{(train)} = (X_t^{(train)}, Y_t^{(train)})$ and a testing split $D_t^{(test)} = (X_t^{(test)}, Y_t^{(test)})$. The chosen splitting method is arbitrary. The training process for each sample will be conducted with $D_t^{(train)}$ and evaluation with $D_t^{(test)}$.
[25] mention different measures for model performance, stability and plasticity. I will focus on the dynamic forms given by [7], because they are adapted for in training use i.e. they represent a model's current state after the t-th training step.
Accuracy $\mathbf{A}$ represents a models performance i.e. how well the predictions $\hat{Y}_t^{(test)}$ align with the true values of $Y_t^{(test)}$ for a metric μ . When $A_{i,k}$ is the accuracy measured on the k-th test split after the i-th training step, then

$$\mathbf{A} = \frac{2}{t(t-1)} \sum_{i \geq k}^{t} A_{i,k} \tag{10}$$

is the average accuracy after the t-th training step over all test splits $D_k^{(test)}, k <= t$ [7].
Backward Transfer $\mathbf{BWT}$ evaluates a models stability [25]. The metric quantifies the influence of learning sample $D_{t+1}^{(train)}$ has on the performance over test sample $D_t^{(test)}$ [16]. Given, the above mentioned, individual *Accuracy* scores $A_{i,k}$

$$\mathbf{BWT} = \frac{2}{t(t-1)} \sum_{i=2}^{t} \sum_{k=1}^{i-1} (A_{i,k} - A_{k,k}) \tag{11}$$

is the average backward transfer after the t-th training step [7]. Note that $\mathbf{BWT}$ can be negative. This property captures (catastrophic) forgetting [25].

Forward Transfer **FWT** is a metric for model plasticity [25]. Complementary to BWT, *Forward Transfer* measures how previous training steps influence the current one. Again the individual *Accuracy* scores are the basis for this evaluation metric. The average influence of old training steps on the model performance after the t-th step:

$$\mathbf{FWT} = \frac{2}{t(t-1)} \sum_{i<k}^{t} A_{i,k} \tag{12}$$

[7].

Another metric that directly measures the relationship between stability and plasticity is presented in [20]. The authors use the maximum eigenvalue of the loss' Hessian λ^{max} to describe the width of their approximation of the loss' minimum. They hypothesize that the *wideness* of this minimum correlates with the forgetting rate of the respective model. Given W_t^* and W_{t+1}^* the optimal parameters after learning the t-th and $t+1$-th task and $L_t(\cdot)$ and $L_{t+1}(\cdot)$ the corresponding loss functions. Mirzadeh et al. formulate the upper bound

$$F_t = L_t(W_{t+1}^*) - L_t(W_t^*) \approx \frac{1}{2}\Delta W^\top \nabla^2 L_t(W_t^*)\Delta W \leq \frac{1}{2}\lambda_t^{max}\|\Delta W\|^2 \tag{13}$$

for the forgetting F_t of the t-th task. They approximate $L_t(W_{t+1}^*)$ around W_t^* with a second order Taylor approximation, where ∇^2 is the Hessian for L_t and ΔW the difference between W_{t+1}^* and W_t^*. They argue that the loss can be approximated this way, because of its almost convex path around the minimum, for models that have more observations per sample than parameters.

Further, ΔW is dependent on the training process of the $t+1$-th task, which depends on the random sample it is trained on, so one can view the differences in parameters as a random vector, that follows some distribution parameterized by the eigenvalues of $\nabla^2 L_t(W_t^*)$ [20].

Controlling the distance of the weights seems to be the key to mitigating forgetting...

5 Regularization

As mentioned in subsection 3.2, one way to address the stability-plasticity problem is the use of regularization. This approach adds a penalty term to the loss function of a model. Usually this penalty term depends on the model parameters. Later we will also see some methods that directly penalize the output of a model. I will begin by categorizing the regularization methods that I have found through out my research and present some

selected examples. After this overview of current possibilities in regularization techniques, I will present attempts at unifying and generalizing this field.

5.1 Regularization via Parameters

Assuming a CL problem with $T = 2$ linear regression tasks. The task corresponding samples $D_1 = (X_1, y_1)$ and $D_2 = (X_2, y_2)$ do not necessarily come from the same population. The *ordinary conitunal learning* [9, 15] algorithm performs an ordinary least square minimization over the first sample set D_1 to estimate the parameters

$$w_1 = (X_1^\top X_1)^{-1} X_1^\top y_1 \tag{14}$$

. In the second training sequence, ordinary continual learning fits D_2 to the residuals of task one with respect to X_2. The new parameters w_2 are then:

$$w_2 = w_1 + (X_2^\top X_2)^{-1} X_2^\top (y_2 - X_2 w_1) \tag{15}$$

. In their analysis of ordinary continual learning [15] show that it suffers from catastrophic forgetting when dealing with "dissimilar" tasks i.e.[15] measure similarity via the following bound:

$$d_F \leq tr(H_1 H_2^{-1}) = o(n) \tag{16}$$

where d_F is the normed expected forgetting rate between the two tasks and $H_i, i \in \{1, 2\}$ are the commutable covariance matrices $\frac{1}{n} X_i^\top X_i$.

Regularization of model weights, in a CL setting, penalizes based on the contribution to previous learning steps [25]. One way of measuring a weights influence is the Fisher information matrix. Kirkpatrick et al. [14] justify this approach through a probabilistic view of neural networks. They no longer want to find the parameters that best fit the data pattern but find the most probable model weights, depending on a given data sample. Using Bayes' Rule and the assumption of independent samples (e.g. CIL), they express the conditional probability $\mathbb{P}(w|\mathcal{D}^{(t)}), \mathcal{D}^{(t)} = \{D_1, ..., D_t\}$ of the weights as

$$\log(\mathbb{P}(w|\mathcal{D}^{(t)})) = \log(\mathbb{P}(D_t|w)) + \log(\mathbb{P}(w|\mathcal{D}^{(t-1)})) - \log(\mathbb{P}(D_t)) \tag{17}$$

6 Conclusion

Blub bla bli

A Appendix

A.1 Expansion of eq. 2 in [14] for T samples

Let $D_i, i \in \{1, ..., t\}$ be t independent samples, as described in section 3, $\mathcal{D}^{(t)} = \{D_1, ..., D_t\}$ the joint samples and $w \in \mathbb{R}^d$ a weight vector. Then the conditional probability

$$\begin{aligned} \mathbb{P}(\mathcal{D}^{(t)}|w) &= \frac{\mathbb{P}(D_1, ..., D_t, w)}{\mathbb{P}(w)} \\ &= \frac{\mathbb{P}(D_1, ..., D_{t-1}|D_t, w)\mathbb{P}(D_t, w)}{\mathbb{P}(w)} \\ &= \mathbb{P}(D_1, ..., D_{t-1}|w)\mathbb{P}(D_t|w) \end{aligned} \tag{18}$$

This we plug into the Bayes' Rule for the posterior $\mathbb{P}(w|\mathcal{D}^{(t)})$ and get

$$\begin{aligned} \mathbb{P}(w|\mathcal{D}^{(t)}) &= \frac{\mathbb{P}(\mathcal{D}^{(t)}|w)\mathbb{P}(w)}{\mathbb{P}(\mathcal{D}^{(t)})} \\ &= \frac{\mathbb{P}(D_1, ..., D_{t-1}|w)\mathbb{P}(D_t|w)\mathbb{P}(w)}{\mathbb{P}(\mathcal{D}^{(t)})} \\ &= \frac{\mathbb{P}(w|D_1, ..., D_{t-1})\mathbb{P}(D_t|w)}{\mathbb{P}(D_t)} \end{aligned} \tag{19}$$

The approximate Gaussian for the posterior $\mathbb{P}(w|D_1, ..., D_{t-1})$ of all prior tasks is then $N(w, (\sum_{i=1}^{t-1} \text{diag}(F_i))^{-1})$ using the chain rule for independent Fisher information $F_i = \mathcal{I}_{D_i}(w)$.

B Electronic appendix

Data, code and figures are provided in electronic form.

References

[1] R. Aljundi, K. Kelchtermans, and T. Tuytelaars. Task-free continual learning, 2019. URL https://arxiv.org/abs/1812.03596.

[2] R. Aljundi, M. Lin, B. Goujaud, and Y. Bengio. Gradient based sample selection for online continual learning, 2019. URL https://arxiv.org/abs/1903.08671.

[3] S. H. Bach and M. A. Maloof. *A Bayesian Approach to Concept Drift*, pages 127–135. 2010.

[4] S. A. Bidaki, A. Mohammadkhah, K. Rezaee, F. Hassani, S. Eskandari, M. Salahi, and M. M. Ghassemi. Online continual learning: A systematic literature review of approaches, challenges, and benchmarks, 2025. URL https://arxiv.org/abs/2501.04897.

[5] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. S. Torr, and M. Ranzato. On tiny episodic memories in continual learning, 2019. URL https://arxiv.org/abs/1902.10486.

[6] K.-L. Du and M. N. S. Swamy. *Neural Networks and Statistical Learning*. Springer London, 2 edition, 2019.

[7] N. Díaz-Rodríguez, V. Lomonaco, D. Filliat, and D. Maltoni. Don't forget, there is more than forgetting: new metrics for continual learning, 2018. URL https://arxiv.org/abs/1810.13166.

[8] S. Ebrahimi, F. Meier, R. Calandra, T. Darrell, and M. Rohrbach. Adversarial continual learning, 2020. URL https://arxiv.org/abs/2003.09553.

[9] I. Evron, E. Moroshko, R. Ward, N. Srebro, and D. Soudry. How catastrophic can catastrophic forgetting be in linear regression?, 2022. URL https://arxiv.org/abs/2205.09588.

[10] L. Fahrmeir, T. Kneib, S. Lang, and B. D. Marx. *Regression - Models, Methods and Applications*. Springer Berlin, 2 edition, 2022.

[11] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra. Pathnet: Evolution channels gradient descent in super neural networks, 2017. URL https://arxiv.org/abs/1701.08734.

[12] R. M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(99)01294-2. URL https://www.sciencedirect.com/science/article/pii/S1364661399012942.

[13] K. Javed and M. White. Meta-learning representations for continual learning, 2019. URL https://arxiv.org/abs/1905.12588.

[14] J. Kirkpatrick, R. Pascanu, N. Rabinowitza, J. Veness, A. A. R. Guillaume Desjardins, K. Milan, J. Quan, T. Ramalho, D. H. Agnieszk Grabska-Barwinska, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *arXiv:1612.00796v2*, 2017.

[15] H. Li, J. Wu, and V. Braverman. Fixed design analysis of regularization-based continual learning, 2024. URL https://arxiv.org/abs/2303.10263.

[16] D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning, 2022. URL https://arxiv.org/abs/1706.08840.

[17] A. Mallya, D. Davis, and S. Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights, 2018. URL https://arxiv.org/abs/1801.06519.

[18] J. Mcclelland, B. Mcnaughton, and R. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102 3:419–457, 1995. doi: https://doi.org/10.1037/0033-295X.102.3.419.

[19] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

[20] S. I. Mirzadeh, M. Farajtabar, R. Pascanu, and H. Ghasemzadeh. Understanding the role of training regimes in continual learning, 2020. URL https://arxiv.org/abs/2006.06958.

[21] R. Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological RReview*, 97 2:285–308, 1990. URL https://api.semanticscholar.org/CorpusID:18556305.

[22] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning, 2017. URL https://arxiv.org/abs/1611.07725.

[23] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learing: From Theory to Algorithms*. Cambridge University Press, 2014.

[24] G. M. van de Ven, T. Tuytelaars, and A. S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, Dec 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00568-3. URL https://doi.org/10.1038/s42256-022-00568-3.

[25] L. Wang, X. Zhang, H. Su, J. Zhu, Fellow, and IEEE. A comprehensive survey of continual learning: Theory and method and application, 2024. URL https://arxiv.org/abs/2302.00487.

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, March 20th, 2025

Name