A Theoretical Analysis of Catastrophic Forgetting

Pierre Alquier

RIKEN

AIP
Center for
Advanced Intelligence Project

New Trends in Statistical Learning II
Porquerolles, June 2022

Thang Doan

Bosch research

Mehdi Abbana Bennani

Aqemia

Bogdan Mazoure

MILA / McGill University

Guillaume Rabusseau

MILA / Unifersité de Montréal

Doan, T., Bennani, M. A., Mazoure, B., Rabusseau, G. & Alquier, P. (2021). A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. *AISTATS'2021*.

Contents

Contents

Notations

Regression/classification problem :

- objects $x \in \mathcal{X}$,
- labels $y \in \mathcal{Y} \subset \mathbb{R}$,
- predictors $f_w : \mathcal{X} \to \mathcal{Y}$, $w \in \mathbb{R}^d$, objective : neural networks.

Difficulties of "continual learning"

- d is huge, $\to$ we need a lot of data.
- the dataset is huge, $\to$ impossible to store all the data.
- we will learn w sequentially based on a data stream (x_t, y_t), $\to$ the x_t come from a real life data collection process that makes them non-indentically distributed..

Online learning theory

Online learning theory provides algorithms to learn from data streams, with theoretical guarantees.

Online Gradient Algorithm

- $w_1 := 0$,
- $w_{t+1} = w_t - \eta_t \nabla_{w=w_t} \ell(y_t, f_w(x_t))$.

Regret bound for OGA

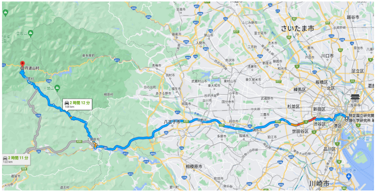
If ℓ is L-Lipschitz + convex, one can calibrate η_t such that

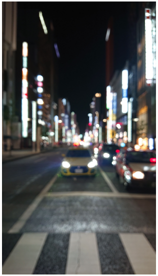
$$\frac{1}{T}\sum_{t=1}^{T} \ell(y_t, f_{w_t}(x_t)) - \inf_{\|w\|\leq B} \frac{1}{T}\sum_{t=1}^{T} \ell(y_t, f_w(x_t)) \leq BL\sqrt{\frac{2}{T}}.$$

Example : training a self-driving car

Decide an itinerary

- from RIKEN AIP (Tokyo)
- to Tabayama.





Observation

$$y_t = f_{w^*}(x_t)$$

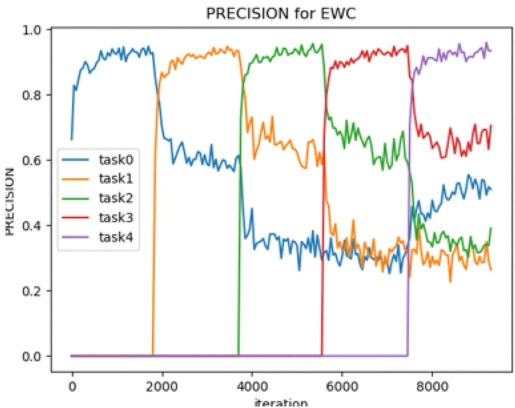
- $y = 1, \dots, \tau_1$:
  - x_t i.i.d from $P_1 \rightarrow$ we learn w_1.
- $y = \tau_1 + 1, \dots, \tau_2$:
  - x_t i.i.d from $P_2 \rightarrow$ we update w_1 to w_2.
- $\dots$
- $y = \tau_K + 1, \dots, \tau_{K+1}$:
  - x_t i.i.d from $P_K \rightarrow$ we update w_K to w_{K+1}.
- $x \sim P_1$:
  - $f_{w_{K+1}}(x)$ is a much worse prediction than $f_{w_1}(x)$.
  - we forgot how to deal with objects $x \sim P_1$.

What is the problem with online learning theory ?

$$\frac{1}{T}\sum_{t=1}^{T}\ell(y_t, f_{w_t}(x_t)) - \inf_{\|w\|\leq B}\frac{1}{T}\sum_{t=1}^{T}\ell(y_t, f_w(x_t)) \leq BL\sqrt{\frac{2}{T}}.$$

- tells you $f_{w_t}(x_t)$ predicts well y_t (on average over t), *not* that $f_{w_T}(x_t)$ predicts well y_t.
- *online-to-batch* bounds : averaging $\bar{w}_t = \frac{1}{t}\sum_{s=1}^{t} w_s$ is proven to work well for out-of-sample prediction... in the i.i.d case !

An example



Hong, D. Y., Li, Y. & Shin, B. S. (2019). Predictive EWC : mitigating catastrophic forgetting of neural network through pre-prediction of learning data. *Journal of Ambient Intelligence and Humanized Computing*.

Some references

Sutton, R. (1986). Two problems with back propagation and other steepest descent learning procedures for networks. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A. & Hassabis, D. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*.

Kemker, R., McClure, M., Abitino, A., Hayes, T. & Kanan, C. (2018). Measuring catastrophic forgetting in neural networks. *AAAI'2018*.

Contents

Linear model – notations

- initialization : $w_{\tau_0} = 0$.
- task τ_k given as a block :

$$\mathrm{Y}_{\tau_k} := \begin{pmatrix} y_{\tau_k+1} \\ \vdots \\ y_{\tau_{k+1}} \end{pmatrix} \text{ and } \mathrm{X}_{\tau_k} := \begin{pmatrix} x_{\tau_k+1}^T \\ \hline \vdots \\ \hline x_{\tau_{k+1}}^T \end{pmatrix}$$

- update :

$$\begin{aligned} w_{\tau_k} &= \arg\min_{w \in \mathbb{R}^d} \left\{ \|\mathrm{Y}_{\tau_k} - \mathrm{X}_{\tau_k} w\|^2 + \lambda . \|w - w_{\tau_{k-1}}\|^2 \right\} \\ &= w_{\tau_{k-1}} + (\mathrm{X}_{\tau_k}^T \mathrm{X}_{\tau_k} + \lambda . \mathrm{I})^{-} \mathrm{X}_{\tau_k}^T \underbrace{\left(\mathrm{Y}_{\tau_k} - \mathrm{X}_{\tau_k} w_{\tau_{k-1}} \right)}_{= \tilde{\mathrm{Y}}_{\tau_k}} . \end{aligned}$$

Definition of forgetting

Definition - forgetting of task *i* at the end of task *j*

For $s \leq t$ we put

$$\Delta^{\tau_s \to \tau_t} := \|X_{\tau_s} w_{\tau_t} - X_{\tau_s} w_{\tau_s}\|^2.$$

- $X_{\tau_t} = U_{\tau_t} \Sigma_{\tau_t} V_{\tau_t}^T$ be the SVD of X_{τ_t},
- $O^{\tau_s \to \tau_t} = V_{\tau_s}^T V_{\tau_t}$ the overlap matrix,
- $M_{\tau_t} := \Sigma_{\tau_t} (\Sigma_{\tau_t} + \lambda.I)^{-} U_{\tau_t}^T$.

Theorem

For any $t > s$,

$$\Delta^{\tau_s \to \tau_t} = \left\| \sum_{k=s+1}^{t} U_{\tau_k} \Sigma_{\tau_k} O^{\tau_s \to \tau_k} M_{\tau_k} \tilde{Y}_{\tau_k} \right\|^2 .$$

Upper bound on forgetting

Corollary

$$\sqrt{\Delta^{\tau_s \to \tau_t}} \leq \|\Sigma_{\tau_s}\|_{\mathrm{op}} \sum_{k=s+1}^{t} \|O^{\tau_s \to \tau_t}\|_{\mathrm{op}} \left\| M_{\tau_k} \tilde{Y}_{\tau_k} \right\|$$

With $V_{\tau_t} = (V_{\tau_t}[1] | V_{\tau_t}[2] | \dots)$ we have

$$O_{i,j}^{\tau_s \to \tau_t} = \cos(V_{\tau_s}[i], V_{\tau_t}[j])$$

and $\|O^{\tau_s \to \tau_t}\|_{\mathrm{op}} = \cos(\alpha)$ where α is the Dixmier angle between the span of V_{τ_t} and the span of V_{τ_s}.

Dixmier, J. (1949). Étude sur les variétés et les opérateurs de Julia, avec quelques applications. *Bulletin de la SMF*.

A recent improvement

Evron, I., Moroshko, E., Ward, R., Srebro, N. & Soudry, D. (2022). How catastrophic can catastrophic forgetting be in linear regression ? *COLT'22*.

- simplified setting, allows an refinement of the analysis,
- note : I find their results very elegant, so I presented the previous result using *some* of their notations.

In their paper :

- $\lambda = 0$, there is w^* such that $\mathrm{Y}_{\tau_s} = \mathrm{X}_{\tau_s} w^*$ (no noise).
- the X_{τ_s} are normalized $\Rightarrow \|\Sigma_{\tau_s}\|_{\mathrm{op}} \leq 1$.

Consequences of the simplifications

Define the orthogonal projection $\mathrm{P}_{\tau_k} = I - \mathrm{X}_{\tau_k}(\mathrm{X}_{\tau_k}^T \mathrm{X}_{\tau_k})^{-} \mathrm{X}_{\tau_k}^T$,

$$\text{then } w_{\tau_k} - w^* = \mathrm{P}_{\tau_k}(w_{\tau_{k-1}} - w^*) = \mathrm{P}_{\tau_k} \dots \mathrm{P}_{\tau_1}(\underbrace{w_{\tau_0}}_{=0} - w^*),$$

$$\begin{aligned} \text{and } \Delta^{\tau_s \to \tau_t} &= \|\mathrm{X}_{\tau_s} w_{\tau_t} - \mathrm{X}_{\tau_s} w_{\tau_s}\|^2 \\ &= \|\mathrm{X}_{\tau_s} w_{\tau_t} - \mathrm{Y}_{\tau_s}\|^2 \\ &= \|\mathrm{X}_{\tau_s} w_{\tau_t} - \mathrm{X}_{\tau_s} w^*\|^2 \\ &= \|\mathrm{X}_{\tau_s} \mathrm{P}_{\tau_t} \dots \mathrm{P}_{\tau_1} w^*\|^2 \\ &\leq \|(\mathrm{I} - \mathrm{P}_{\tau_s}) \mathrm{P}_{\tau_t} \dots \mathrm{P}_{\tau_1} w^*\|^2. \end{aligned}$$

Average forgetting : worst case

Definition - average forgetting at task t

$$F(t) := \frac{1}{t}\sum_{s=1}^{t} \|X_{\tau_s} w_{\tau_t} - X_{\tau_s} w_{\tau_s}\|^2 = \frac{1}{t}\sum_{s=1}^{t} \Delta^{\tau_s \to \tau_t}$$

$$F(t) = \frac{1}{t}\sum_{s=1}^{t} \|X_{\tau_s} P_{\tau_t} \dots P_{\tau_1} w^{*}\|^2$$

They design a situation where :

$$F(t) \geq 1 - \mathcal{O}\left(\frac{1}{\sqrt{t}}\right).$$

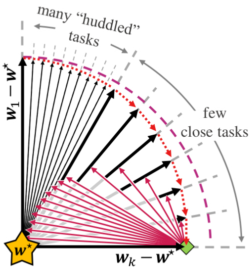


Figure from Evron *et al.* (2022).

Situations where forgetting do not occur

Evron *et al.* (2022) then argue that in general, forgetting is not that bad :

- cyclic tasks : $\tau_1, \dots, \tau_T, \tau_1, \dots, \tau_T, \dots$ After seeing t tasks,

$$F(t) \leq \min\left(\frac{T^2}{\sqrt{t}}, \frac{T^2(d - \max\{\mathrm{rank}(X_{\tau_s})\}}{t} \right),$$

- randomized tasks : $\tau_{I_1}, \tau_{I_2}, \dots$ where the I_i are i.i.d uniform in $\{1, \dots, T\}$, then after seeing t tasks,

$$\mathbb{E}[F(t)] \leq \frac{9\left(d - \frac{1}{T}\sum_{s=1}^{T} \mathrm{rank}(X_{\tau_s})\right)}{t}.$$

→ however, this requires to store the tasks, or, at least, to be able to learn them many times...

Conclusion of the theoretical analysis

What we learnt so far

- catastrophic forgetting can happen even in linear models,
- depends on the geometry and order of the tasks.

Open questions :

- noisy case,
- nonlinear case,
- tasks not by block // not aware that a new task begins,
- other algorithms... (we propose a few in the next section),
- theoretical limitations :

Knoblauch, J., Hisham, H. & Diethe, T. (2020). Optimal continual learning has perfect memory and is NP-hard. *ICML'2020*.

Contents

Orthogonal updates

Doan, T., Bennani, M. A. & Sugiyama, M. (2020). Generalisation guarantees for continual learning with orthogonal gradient descent. *ICML'2020 Workshop on Lifelong Learning*.

$$w_{\tau_k} = \underset{\substack{w \in \mathbb{R}^d \\ V_{\tau_1}^T (w - w_{\tau_{k-1}}) = 0 \\ \vdots \\ V_{\tau_{k-1}}^T (w - w_{\tau_{k-1}}) = 0}}{\arg\min} \left\{ \|Y_{\tau_k} - X_{\tau_k} w\|^2 + \lambda . \|w - w_{\tau_{k-1}}\|^2 \right\}$$

$$= w_{\tau_{k-1}} + \Pi_k (X_{\tau_k}^T X_{\tau_k} + \lambda . I)^{-} X_{\tau_k}^T \left(Y_{\tau_k} - X_{\tau_k} w_{\tau_{k-1}} \right)$$

where Π_k is the orthogonal projection on $\ker(V_{\tau_1}^T | \dots | V_{\tau_{k-1}}^T)$.

$$\Delta^{\tau_s \to \tau_t} = 0.$$

But the procedure requires to store V_{τ_1}, V_{τ_2}, ...

Data compression (1/2)

In general, $\underbrace{\mathrm{X}_{\tau_t}}_{N_t \times d} = \underbrace{\mathrm{U}_{\tau_t}}_{N_t \times N_t} \underbrace{\Sigma_{\tau_t}}_{N_t \times N_t} \underbrace{\mathrm{V}_{\tau_t}^T}_{N_t \times d}$.

Data compression : replace V_{τ_t} by $\hat{\mathrm{V}}_{\tau_t}$ $(d \times n,\ n \ll N_t)$:

- "OGD" : $\hat{\mathrm{X}}_{\tau_t}$: n rows sampled from X_{τ_t}, $\hat{\mathrm{X}}_{\tau_t} = \hat{\mathrm{U}}_{\tau_t} \hat{\Sigma}_{\tau_t} \hat{\mathrm{V}}_{\tau_t}^T$.

Farajtabar, M., Azizan, N., Mott, A. & Li, A. (2020). Orthogonal gradient descent for continual learning. *AISTATS'2020*.

- instead of random rows, "memorable observations" :

Pan, P. , Swaroop, S. , Immer, A., Eschenhagen, R., Turner, R. & Khan, M. E. (2020). Continual Deep Learning by Functional Regularisation of Memorable Past. *NeurIPS'2020*.

Different framework, but the philosophy would here lead to select high-leverage observations.

Data compression (2/2)

Data compression : replace V_{τ_t} by $\hat{V}_{\tau_t}$ $(n \times d,\ n \ll N_t)$:

- our proposal, "PCA-OGD" : PCA on X_{τ_t}, that is

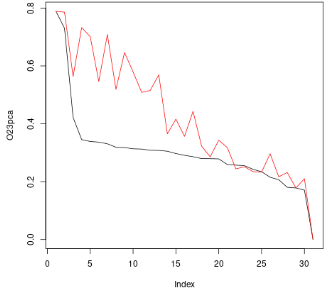
$$V_{\tau_t}^T = \left(\frac{\hat{V}_{\tau_t}^T}{*} \right).$$

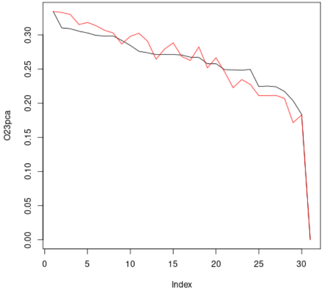
- $\hat{\Pi}_t :=$ orthogonal projection on $\ker(\hat{V}_{\tau_1}^T | \dots | \hat{V}_{\tau_{t-1}}^T)$.
- $\hat{O}^{\tau_s \to \tau_t} = V_{\tau_s}^T \hat{\Pi}_t V_{\tau_t}$

$$\sqrt{\Delta^{\tau_s \to \tau_t}} \leq \|\Sigma_{\tau_s}\|_{\mathrm{op}} \sum_{k=s+1}^{t} \left\| \hat{O}^{\tau_s \to \tau_t} \right\|_{\mathrm{op}} \left\| M_{\tau_k} \tilde{Y}_{\tau_k} \right\|$$

Simulation

$\|\hat{O}^{\tau_s \to \tau_t}\|_{\mathrm{op}}$ for "OGD" and "PCA-OGD" in two settings.





Experiments on the MNIST dataset

Neural network with the NTK approximation :

$$f_w(x) \simeq f_{w_0}(x) + \langle \nabla_{w=w_0} f_w(x), w - w_0 \rangle$$

Experiments : impact of $\|O^{\tau_s \to \tau_t}\|_{\mathrm{op}}$

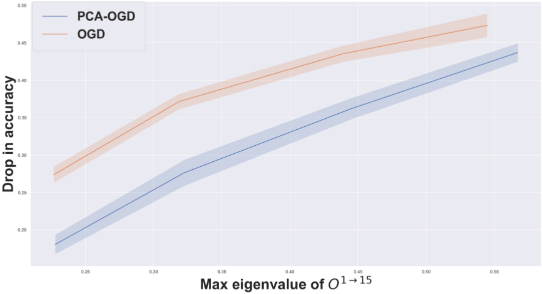
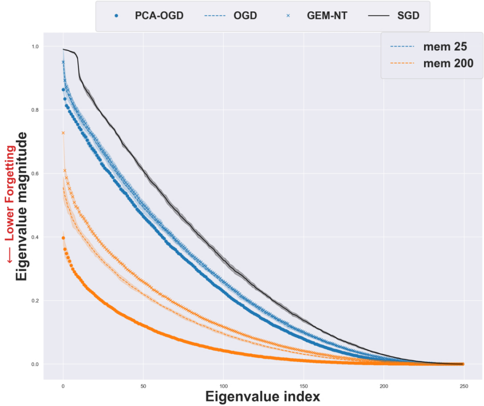


Figure 2: Drop in performance with respect to the maximum eigenvalue for Rotated MNIST (averaged over 5 seeds ± 1 std).

Experiments : evaluation of $\|\hat{O}^{\tau_s \to \tau_t}\|_{op}$



Experiments : performances

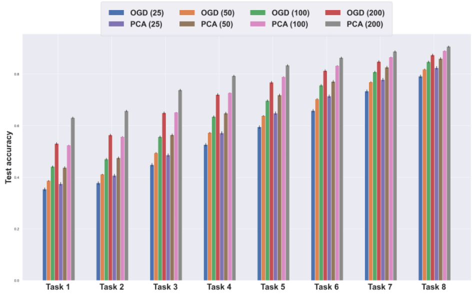


Figure 4: Final accuracy on **Rotated MNIST** for different memory size (averaged over 5 seeds ± 1 std). OGD needs twice as much memory as PCA-OGD in order to achieve the same performance (i.e compare OGD (200) and PCA (100).

La fin

終わり

ありがとう ございます。