

Catapults in SGD

Étude et reproduction expérimentale de *Catapults in SGD*

Valentin Dugay, Jibril El Hassani | ¹CentraleSupélec

Introduction

Ce poster présente une étude du phénomène de *catapultes* : des pics transitoires de loss observés lors de l'entraînement par descente de gradient stochastique (SGD). Nous reproduisons ces effets sur MNIST et évaluons leur impact sur la performance (notamment la généralisation), avec un focus sur l'extension (plus délicate) à AdaGrad.

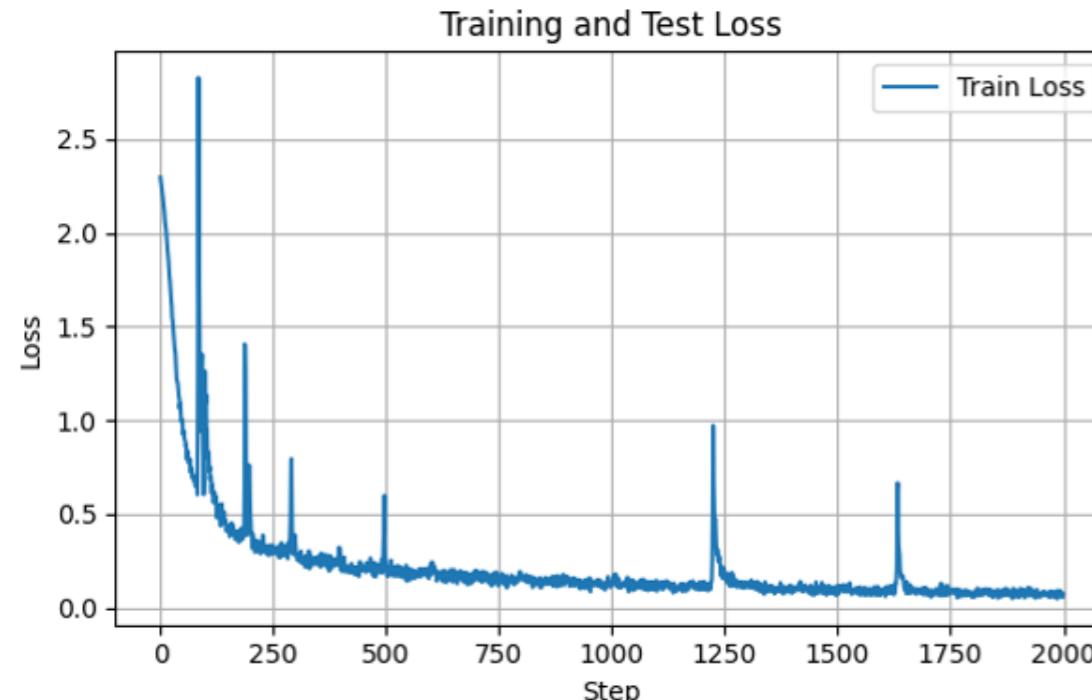


Figure: Illustration des catapultes : pics de loss suivis d'un retour à une dynamique d'entraînement classique.

Origine du phénomène et reproduction

D'après le *descent lemma* :

$$\mathcal{L}(\mathbf{w}^{t+1}) \leq \mathcal{L}(\mathbf{w}^t) - \eta \left(1 - \frac{\eta\beta}{2}\right) \|\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^t)\|^2,$$

avec β la plus grande valeur propre de la Hessienne. Si $\eta > \frac{2}{\beta}$, la loss peut augmenter, typiquement sur un petit nombre de directions associées aux plus grandes valeurs propres : on observe alors des *catapultes* (pics sans divergence). Pour provoquer ces événements de manière contrôlée, nous utilisons des schedules de learning rate (cyclique et cyclique exponentiel) :

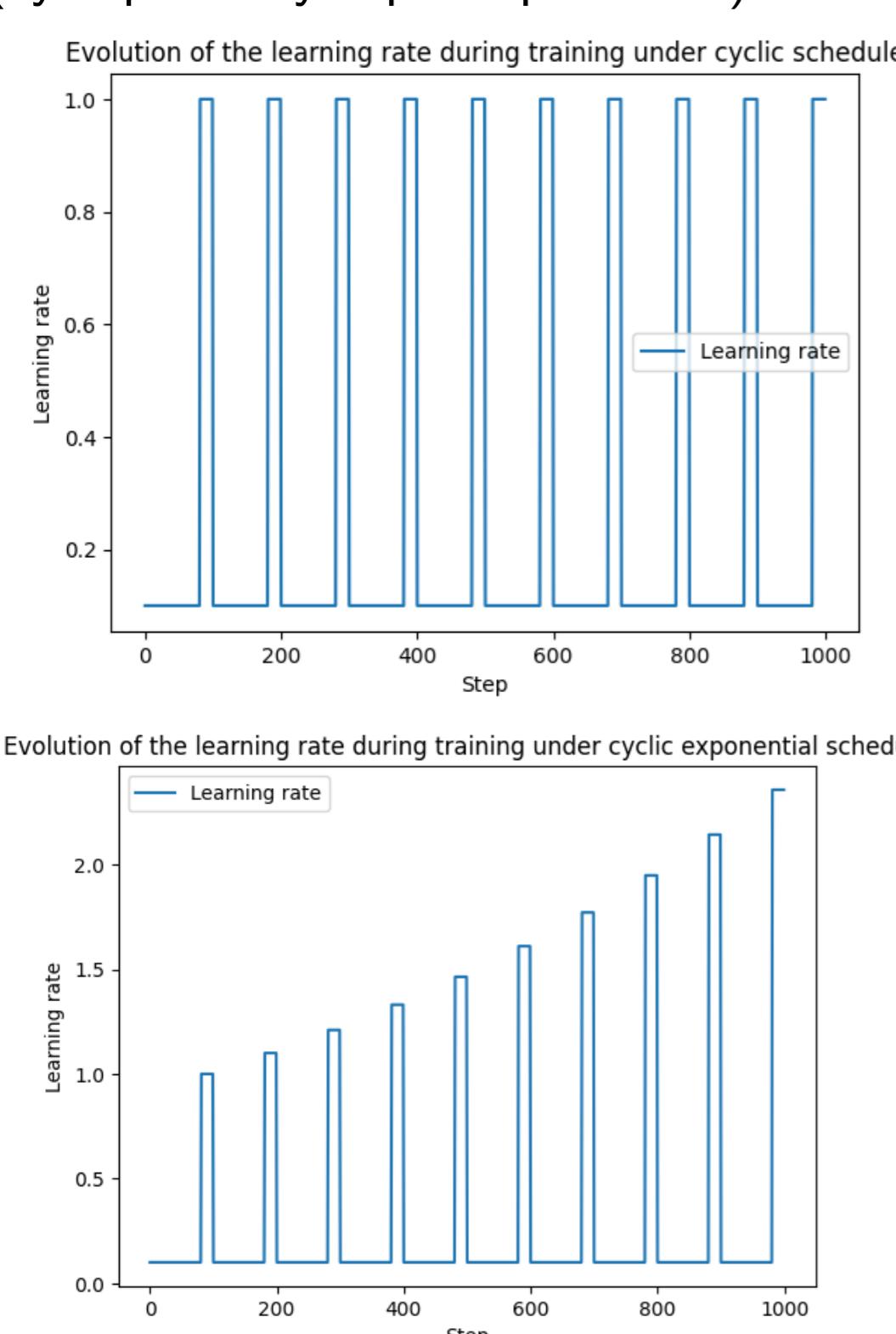


Figure: Schedules de learning rate utilisés : cyclique et cyclique exponentiel.

Catapultes et généralisation (MNIST)

Nous comparons, à protocole d'entraînement identique, plusieurs optimiseurs. Les catapultes améliorent nettement la généralisation de SGD, tandis qu'elles sont plus difficiles à exploiter avec AdaGrad.

Méthode d'optimisation	Test accuracy (%)
SGD	95.38
SGD + Catapult	97.65
AdaGrad	97.09
AdaGrad + Catapult	95.91
Adam	98.12

Table: Comparaison des performances sur MNIST (réseau fully-connected à 2 couches).

Analyse NTK : conditionnement et norme spectrale

En complément, nous analysons la dynamique dans le régime NTK en utilisant un modèle plus petit. Nous suivons le *condition number* et la norme spectrale $\|K\|_2 = \lambda_{\max}(K)$ de la matrice NTK, et visualisons les phases de catapultes (lignes rouges).

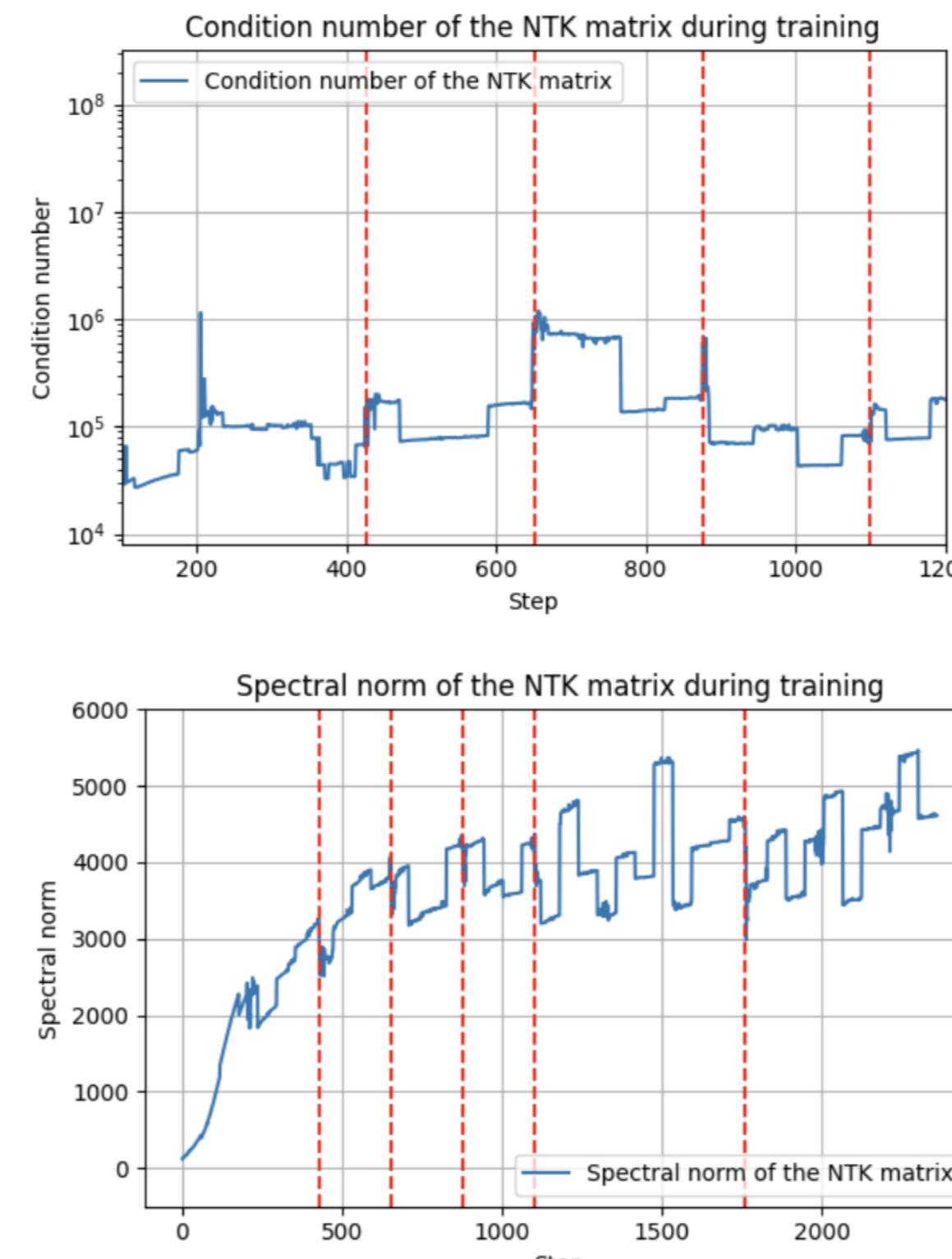


Figure: Condition number et norme spectrale de la matrice NTK, associées phases de catapultes.

Les catapultes sont plus dures à obtenir sur Adagrad que sur SGD

On se propose de montrer que le learning rate critique d'Adagrad $\eta_{crit,ADAGRAD}$ tend vers l'infini quand le nombre d'itérations tend vers l'infini, de plus pour n'importe quel learning rate borné il existe un nombre de pas d'optimisation à partir duquel la loss ne peut plus avoir de catapultes. On se place dans le régime NTK, la descente de gradient stochastique a pour update :

$$f^{t+1} - y = (I - 2\eta \frac{K}{b})(f^t - y)$$

Utiliser Adagrad au lieu de SGD revient à changer le learning rate en fonction du paramètre, sans Adagrad, le learning rate est le même pour tous les paramètres : $(\eta)_i \leq n$. Avec Adagrad, on a pour learning rate :

$$\left(\frac{\eta}{[\sqrt{g} \circ g]}_i\right)_{i \leq n}$$

Où le vecteur $\sqrt{\sum_p g_p \circ g_p}$ est l'historique des carré des gradients pour chaque poid du modèle.

Comme on sait qu'il y a une catapulte si $\eta > \eta_{crit}$, une manière naïve d'analyser les catapultes sur Adagrad est de regarder :

$$\left(\frac{\eta}{s_t}\right)_{i \leq n}$$

avec

$$s_t = \min(\sum_p \sqrt{g_p \circ g_p})$$

. On sait que si ce learning rate reste en dessous du learning rate critique, il n'y aura pas de catapultes.

Étant donné que $\eta_{crit,SGD} = \frac{2}{\lambda_{\max}}$, on a :

$$\eta_{crit,ADAGRAD}^t \geq \frac{2s_t}{\lambda_{\max}^t}$$

Ainsi, en supposant la suite des $(\lambda_{\max}^t)_{t \in \mathbb{N}^*}$ bornée et en supposant $s_t \rightarrow +\infty$ quand $t \rightarrow +\infty$, on a les deux affirmations ci dessus. La Figure 5 montre la validité de nos hypothèses.

Étude pratique des catapultes sur Adagrad

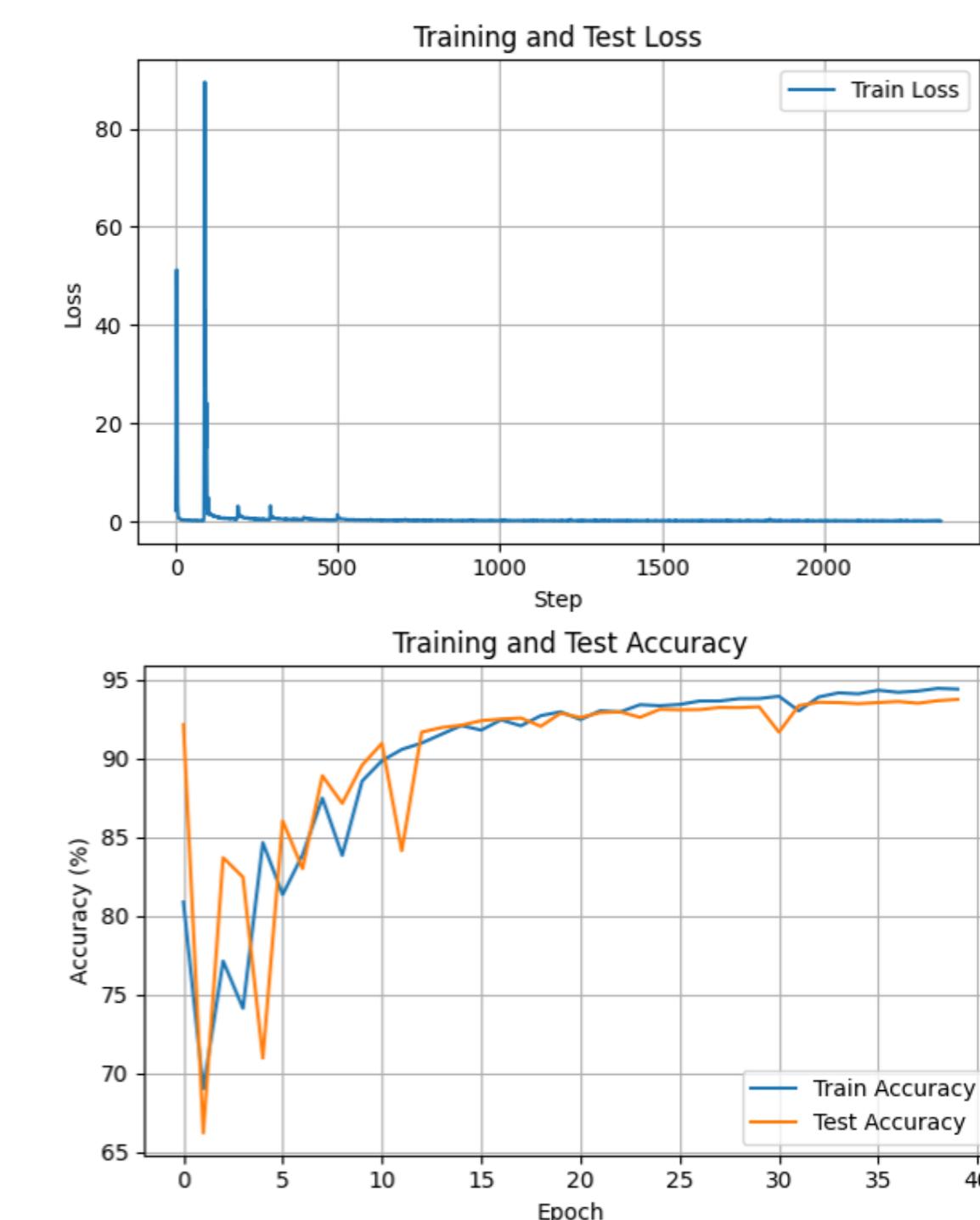


Figure: Évolution de la train loss et de l'accuracy pendant un entraînement utilisant Adagrad avec un scheduler cyclique : on remarque qu'au début, les catapultes sont très fortes et créent donc de l'instabilité (ce qui correspond au début de l'entraînement où l'historique des gradients ont une faible norme), puis que les catapultes disparaissent complètement, car l'historique des gradients tend vers l'infini.

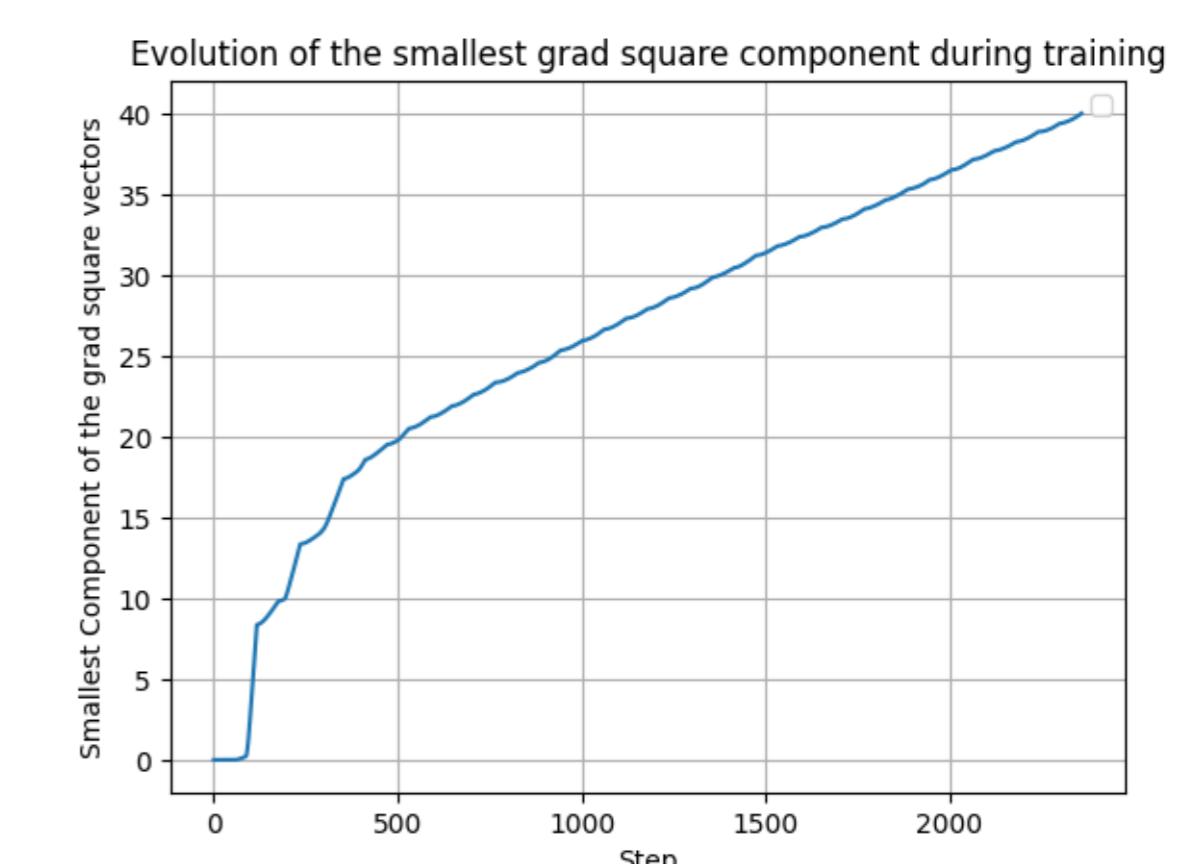


Figure: Évolution de la plus petite composante de la somme cumulative du gradient en fonction du step. On voit que ce terme semble tendre vers l'infini ce qui valide notre hypothèse.

Conclusion

Nous avons reproduit le phénomène de catapultes et mis en évidence son effet bénéfique sur la généralisation de SGD, tout en montrant qu'il est nettement plus difficile de maintenir des catapultes avec AdaGrad. Nous complétons l'étude par un suivi de quantités NTK (conditionnement et norme spectrale) pour relier les phases de catapultes à la dynamique d'entraînement. Les détails expérimentaux sont présentés dans le notebook et le rapport.

Références

- L. Zhu, C. Liu, A. Radhakrishnan, and M. Belkin, *Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning*, arXiv:2306.04815, 2023. <https://arxiv.org/abs/2306.04815>