Sharpness-Aware Optimization for Stability Gap Reduction

Ksenia Sycheva [1] Gido van de Ven [1] Tom Viering [1]

[1]TU Delft

Introduction

In continual learning, deep learning models are trained on multiple tasks sequentially. This approach is useful in many real-world scenarios, but it faces two main challenges:

1. **Catastrophic forgetting**: model might forget tasks that it was trained on before.
2. **Stability gap**: even if catastrophic forgetting does not occur, performance on the previous tasks can drop significantly and then be recovered, which is not efficient. Additionally, this can be critical in safety-related scenarios.

Another research direction explores sharpness-aware optimization for continual learning. These methods optimize neural networks to converge to flat minima (see Figure 1) - regions in the loss landscape known to improve generalization. Recent studies have shown that flat minima can also help mitigate catastrophic forgetting in continual learning.

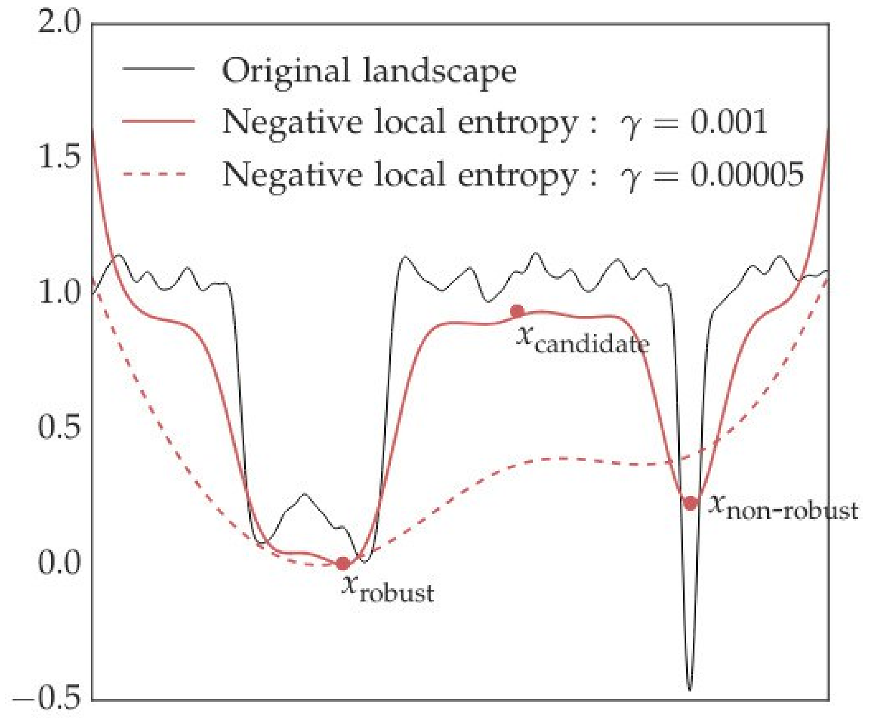

Figure 1. **Entropy-SGD** [2] utilizes local entropy, which concentrates on wide valleys in the energy landscape. Instead of computing loss in a single point, average loss values within neighborhood of this point is approximated via stochastic gradient Langevin dynamics (SGLD) [4]. γ controls the effect of sharpness information by penalizing distance from center point. As $\gamma \to \infty$, standard optimization algorithms are recovered. $\gamma \to 0$ gives uniform loss.
Flat minima can be characterized by Hessian that have most of its eigenvalues close to zero. **C-Flat** [1] in addition to zero-order information computes gradients to approximate second-order curvature properties, which are controlled by hyperparameter ϕ. This way explicitly targets eigenvalues of Hessian, forcing them to be small.

We analyze how sharpness-aware optimization impacts training dynamics in CL, specifically after task transitions. Additionally, we collect empirical evidence that second-order curvature information gives greater control over stability gap.

Research Questions

- **Q1:** Does sharpness-aware optimization contribute to stability gap reduction in continual learning systems?
- **Q2:** Does incorporating second-order information into sharpness-aware optimizers yield additional improvements in stability mitigation?

Methodology

Baselines: considered sharpness-aware methods are optimizer-agnostic (can be applied on top of any optimizer). As baselines we chose two standard optimizers - SGD and Adam - and evaluated their Entropy-regularized and C-Flat variants against these baselines.
Dataset: in all experiments rotated MNIST dataset was used, with three rotation angles seen in fixed order (0°→ 160°→ 80°). Training on each task lasted for 1000 iterations.
Metrics: to evaluate changes in stability gap we calculate maximum decrease in accuracy **MD** after switching tasks, and number of iterations until recovering performance **RS**. In addition to stability gap specific metrics, we compute eventual accuracy on every task to ensure that performance on other tasks is not sacrificed.

Analysis

We analyze Entropy-SGD and C-Flat effect on stability gap. In experiments with both optimizers (SGD and Adam), we can see that training dynamics is affected by incorporating sharpness-aware regularization. Improvement with C-Flat is more consistent, which is likely due to the explicit usage of second-order information in this method.

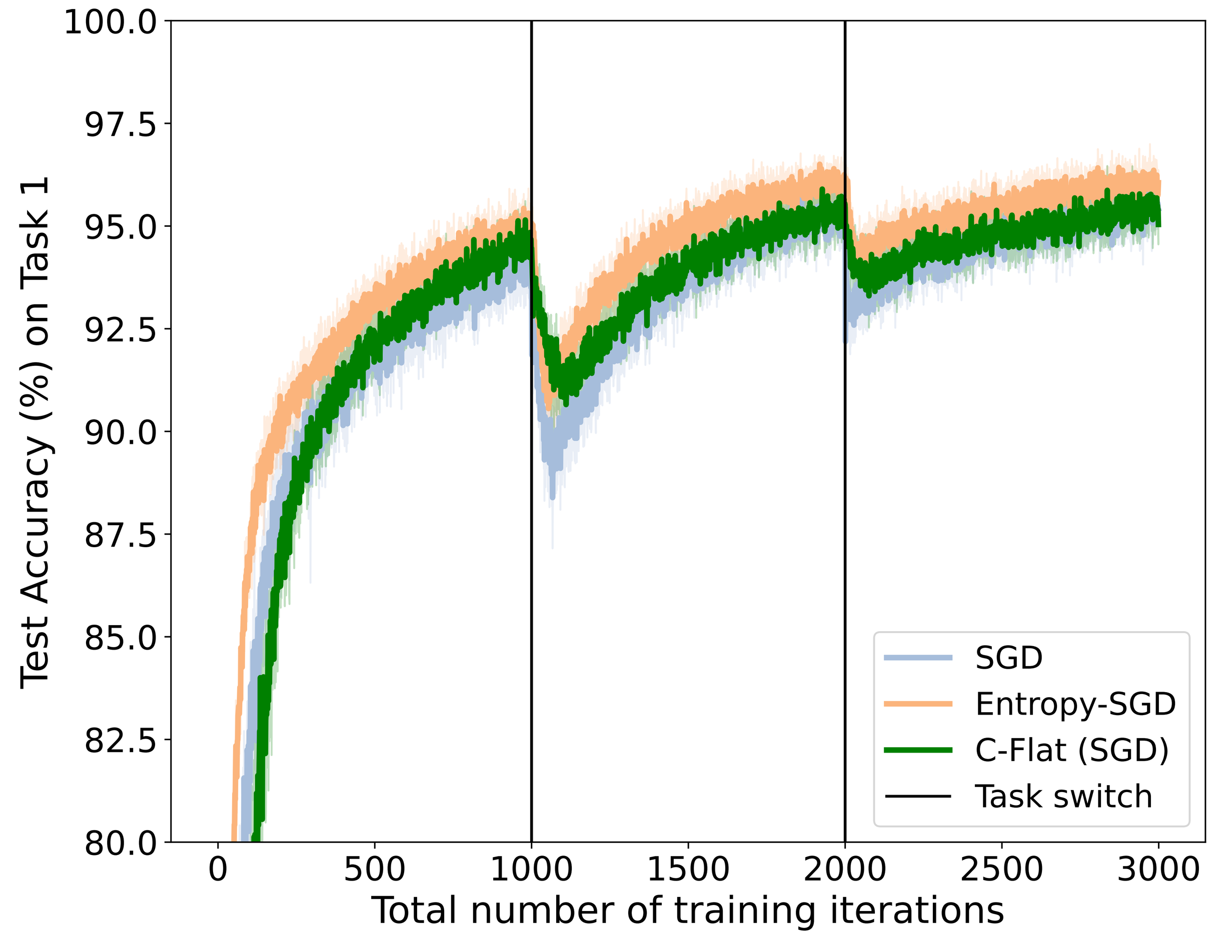

Figure 2. Task 1 accuracy trajectories demonstrating stability gap characteristics of SGD-based optimizers during incremental training (shaded regions indicate ±1 standard error across runs). Both Entropy-SGD and C-Flat exhibit faster recovery from post-switch accuracy drops and better stability preservation compared to vanilla SGD, while simultaneously maintaining competitive downstream task performance.

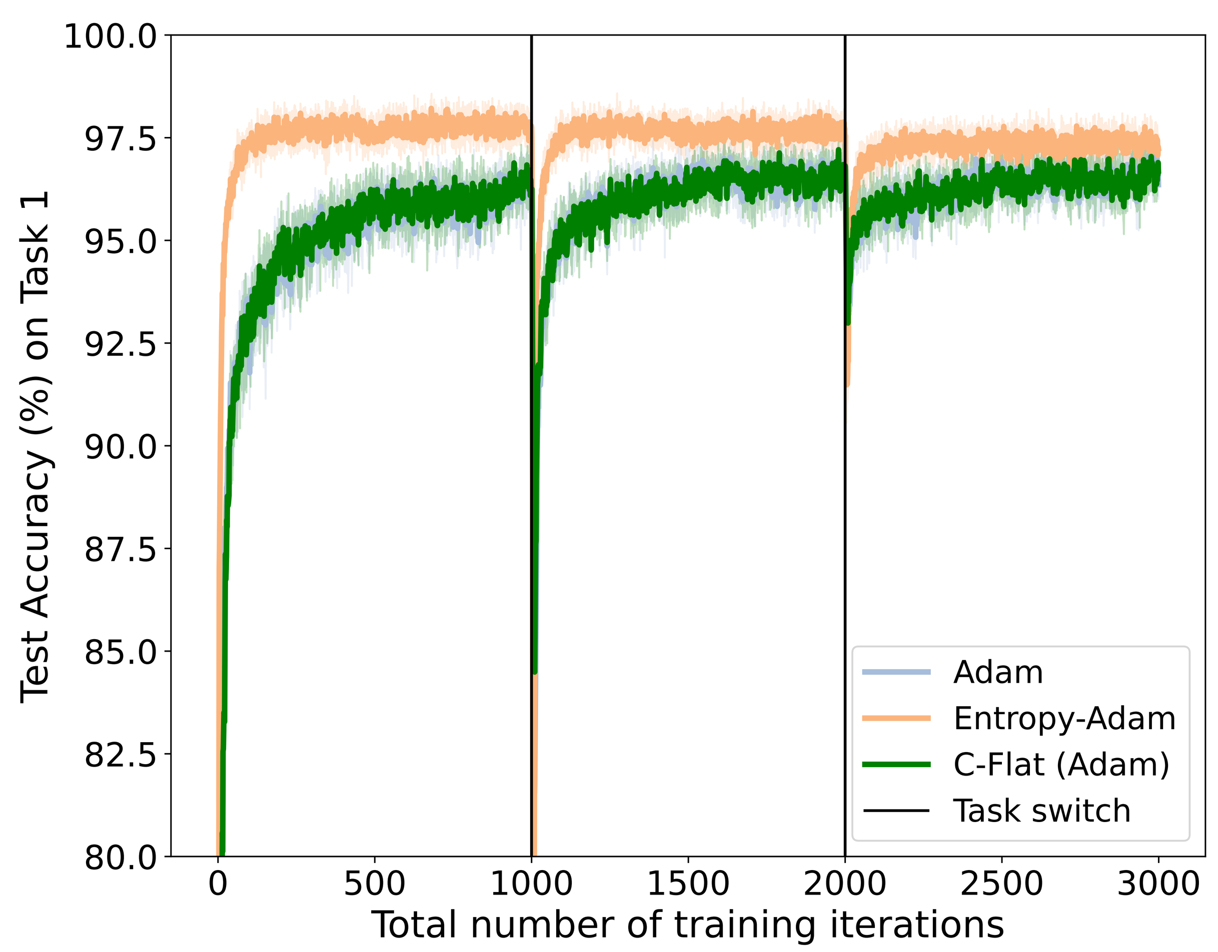

Figure 3. Task 1 accuracy trajectories for Adam-based optimizers, revealing distinct stability gap behaviors. While C-Flat demonstrates faster recovery from task switches, Entropy-Adam shows notably degraded performance compared to both its SGD counterpart and baseline Adam, suggesting the entropy regularization approach may be less suitable in continual learning setting.

Second-Order Information in Optimization

Hessian of flat minima has distinct property: flat minima have most of its eigenvalues with low magnitude. In Entropy-regularized optimization this property is not taken into account explicitly: training is regularized by averaging loss values in the neighborhood around current weights. In contrast to this, C-Flat targets this property directly by using gradients to regularize training objective:

$$\rho \cdot \max\{||\nabla \mathcal{L}(\theta')|| : \theta' \in B(\theta, \rho)\} \tag{1}$$

where $B(\theta, \rho)$ is a ball centered at θ with radius ρ.

Second-Order Analysis

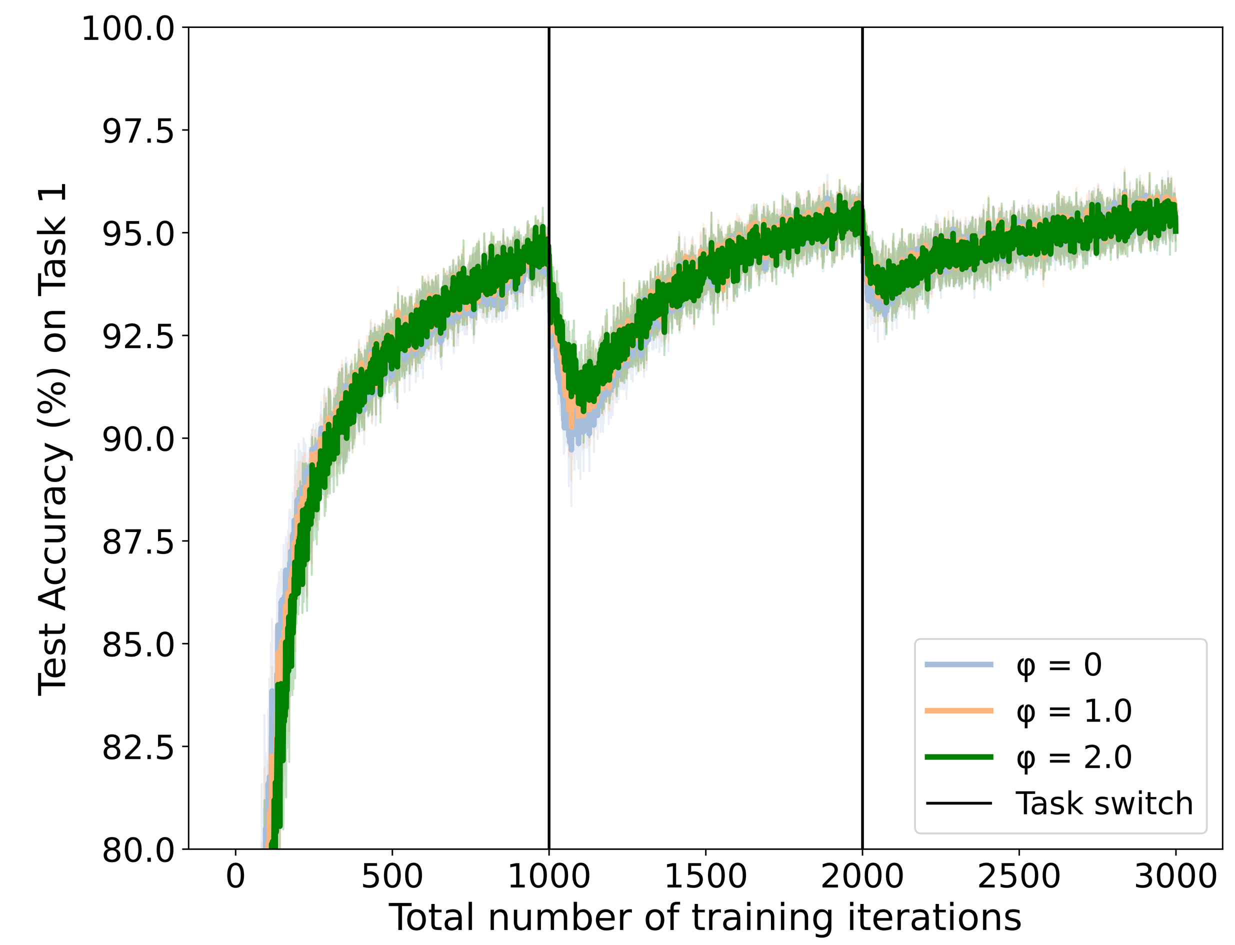

Figure 4. Task 1 accuracy trajectories of C-Flat (SGD) optimizers with different ϕ values during incremental training with different. Larger ϕ values reduce MD more without sacrificing performance significantly.

Conclusion

Overall, sharpness-aware optimization effectively reduces stability gaps, with second-order methods delivering more consistent improvements, without affecting negatively model's performance on other tasks. In future, we want to evaluate sharpness-aware optimizers on longer tasks sequences and test other existing sharpness-aware methods.

References

[1] Ang Bian, Wei Li, Hangjie Yuan, Chengrong Yu, Mang Wang, Zixiang Zhao, Aojun Lu, Pengliang Ji, and Tao Feng. Make continual learning stronger via c-flat, 2024.

[2] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys, 2017.

[3] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization, 2021.

[4] Max Welling and Yee Teh. Bayesian learning via stochastic gradient langevin dynamics. pages 681–688, 01 2011.