Improving Sharpness-Aware Minimisation

Candidate Number:
Mathematical Institute
University of Oxford

Abstract

Modern deep neural networks (DNNs) tend to be over-parameterised leading to overfitting and a gap in generalisation ability. Recently a line of research based on sharpness-aware minimisation (SAM) has shown impressive results by seeking minima with a flat loss landscape within a small neighbourhood. In this report we look at two extensions of this research, surrogate gap minimisation and adaptive sharpness. We then propose a novel optimisation algorithm combining these results, leading to superior performance in image classification tasks. Code can be found at the following anonymous GitHub repository:

https://github.com/a1b2jpg/Theories-of-Deep-Learning/

1 Introduction

Modern DNNs achieve state-of-the-art performance in a wide range of tasks such as image classification, natural language processing, and predictive modelling. However, these models are often over-parameterised, which can lead to overfitting. This can cause a generalisation gap, where the validation accuracy is lower than the training accuracy because the model has learnt the training data rather than the underlying patterns.

SAM addresses this issue by simultaneously minimising both the loss and its sensitivity to parameter perturbations (loss sharpness) within a small neighbourhood. It does this by minimising a quantity known as the perturbed loss.

However, recent research has shown that low perturbed loss values can occur for both flat and sharp minima, limiting SAM's effectiveness [16]. To address this Zhuang et al. introduce the surrogate gap metric, which has a stronger correlation with the flatness of minima. They then propose 'Surrogate Gap Guided Sharpness Aware Minimisation' (GSAM), an algorithm which minimises both loss and surrogate gap, leading to stronger performance of SAM.

A different improvement on SAM is Adaptive sharpness-Aware Minimisation (ASAM). This also refines the approach of SAM by introducing a scale invariant sharpness metric, improving its correlation with generalisation over SAM.

In this paper we propose Adaptive-GSAM, a novel (to the best of our knowledge) optimisation algorithm which combines the scale invariance of ASAM with the surrogate gap measure of GSAM. We show, through a number of numerical experiments that Adaptive-GSAM achieves superior results compared to GSAM on image classification tasks.

2 Why do we want flat minima?

Consider a model $f : X \to Y$ paramaterised by a weight vector $\mathbf{w}$ and a loss function $l : Y \times Y \to \mathbb{R}_+$. Given a sample set of data $S = \{(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n)$, drawn i.i.d from a distribution D, we define the training loss as $L_{\mathcal{S}}(\mathbf{w}) = \sum_{i=1}^{n} l(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w})/n$. Then the generalisation gap between the

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

expected loss $L_{\mathcal{D}}(\mathbf{w}) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l(\mathbf{y}, f(\mathbf{x};\mathbf{w})\right]$ Represents the ability of the model to generalise to unseen data.

Over-parameterised DNNs exhibit complex, non-convex training loss landscapes, characterised by numerous local minima with varying generalization capabilities [11]. The connection between the geometry of the loss landscape and its impact on generalisation has been extensively studied from both theoretical and empirical perspectives [11], [15]. There have been a couple of attempts at using this promising connection to create new optimisation approaches such as in [3] and [9] however there have been few practical applications of these results.

Research into the relationship between generalisation ability and the loss landscape surrounding a global minimum has found that solutions that generalise well tend to lie in flat valleys rather than sharp ravines [8], [7]. Recently, Jiang et al. Conducted an empirical study into 40 complexity measures and found that sharpness had the highest correlation with generalisation[10]. Therefore in order to improve generalisation, we are motivated to find minima within a flat valley.

In this report we measure the sharpness of a local minimum with loss $f(\mathbf{w})$ by the dominant eigenvalue σ_{max} of the Hessian.

3 Sharpness Aware Minimisation

In this section we describe the theoretical foundations of SAM. We start with a PAC-Bayesian generalisation bound, we then demonstrate how this relates to sharpness and outline how this relationship forms the basis for the SAM algorithm.

There have been a number of attempts at penalising sharpness during training, for example by regularising a notion related to minimum descent length or by regularising local entropy. However, these measures are difficult to calculate and differentiate through. Foret et al. Propose SAM by minimising the perturbed loss, a computationally efficient way of penalising sharpness [6]. They begin by bounding generalisation ability by the neighbourhood-wide training loss:

Theorem 1 *With high probability over the set $\mathcal{S}$ generated from distribution $\mathcal{D}$, we have:*

$$L_{\mathcal{D}}(\mathbf{w}) \leq \max_{\|\epsilon\|_2\leq\rho} L_{\mathcal{S}}(\mathbf{w}+\epsilon) + h\left(|\mathbf{w}\|_2^2/\rho^2\right).$$

Where $h : \mathbb{R} \to \mathbb{R}_+$ is a strictly increasing function and $\rho \geq 0$.

We can rearrange the right hand side to make the sharpness term more explicit:

$$[\max_{\|\epsilon\|_2\leq\rho} L_{\mathcal{S}}(\mathbf{w}+\epsilon) - L_{\mathcal{S}}(\mathbf{w})] + L_{\mathcal{S}}(\mathbf{w}) + h\left(\|\mathbf{w}\|_2^2/\rho^2\right).$$

The term in the square brackets captures the sharpness of $L_{\mathcal{S}}$ at $\mathbf{w}$ by measuring the maximum difference in training value within a neighbourhood of size ρ. We choose h to be $\lambda\|\mathbf{w}\|_2^2$ for some hyper-parameter λ.

We can then formulate the sharpness aware minimisation as the following minimax optimisation:

$$\min_{\mathbf{w}} \max_{\|\epsilon\|_p\leq\rho} L_{\mathcal{S}}(\mathbf{w}+\epsilon) + \lambda\|\mathbf{w}\|_{\mathbf{2}}^{\mathbf{2}}$$

SAM solves this by iteratively applying the following two step algorithm for $t = 0, 1, 2...$:

$$\begin{cases} \epsilon_t = \rho\frac{\nabla L_{\mathcal{S}}(\mathbf{w}_t)}{\|\nabla L_{\mathcal{S}}(\mathbf{w}_t)\|_2} \\ \mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t(\nabla L_{\mathcal{S}}(\mathbf{w}_t + \epsilon_{\mathbf{t}}) + \lambda\mathbf{w}_t) \end{cases}$$

SAM estimates the point $\mathbf{w}_t + \epsilon_t$ at which loss is approximately maximised within a neighbourhood of size ρ. It then performs gradient descent at $\mathbf{w}_t$ using the gradient calculated at $\mathbf{w}_t + \epsilon_t$

Whilst SAM produces improved results compared to the base optimiser SGD, It is important to note that SAM requires roughly twice the computations per iteration as SGD as it takes two forward-backward passes per step.

In the next two sections we look at two improvements made on SAM. The first, GSAM, uses a different sharpness measure which has better correlation with the actual sharpness. The second, Adaptive SAM, addresses the scale dependency issue of SAM.

4 Surrogate gap minimisation

In this section we describe the surrogate gap metric, we show that it has a stronger correlation with sharpness than perturbed loss and then describe an algorithm utilising the surrogate gap to achieve improved generalisation results over SAM. We first introduce a couple useful definitions.

Definition 4.1 $\mathbf{w}_t^{adv} := \mathbf{w}_t + \rho_t \frac{\nabla f(\mathbf{w_t})}{\|\nabla f(\mathbf{w_t})\|+\epsilon}$*: equivalent to* $\max_{\|\mathbf{w}'-\mathbf{w}_t\| \leq \rho_t} f(\mathbf{w}')$ *when* ρ_t *is small*

Definition 4.2 *We define the perturbed loss* $f_p(\mathbf{w})$ *as* $f_p := \max_{\|\epsilon\|_2 \leq \rho} L_{\mathcal{S}}(\mathbf{w} + \epsilon)$

Definition 4.3 *We define the surrogate gap* $h(\mathbf{w}) := f_p(\mathbf{w}) - f(\mathbf{w}) \approx f(\mathbf{w}^{adv}) - f(\mathbf{w})$

We can re-characterise SAM as approximately minimising the perturbed loss. We will now show the limitations of this approach and see how to overcome this.

4.1 Perturbed loss is not enough

Contrary to the results found by Foret et al. [6], Zhuang et al. find the following lemma suggesting that the perturbed loss does not necessarily have a strong correlation with sharpness [16].

lemma 1 *For some fixed* ρ*, consider two local minima* $\mathbf{w}_1$ *and* $\mathbf{w}_2$*. We have:*

$$f_p(\mathbf{w}_1) \leq f_p(\mathbf{w}_2) \quad \not\Rightarrow \quad \sigma_{max}(\mathbf{w}_1) \leq \sigma_{max}(\mathbf{w}_2),$$

where σ_{max} *is the dominant eigenvalue of the Hessian.*

They also derive the following two lemmas regarding properties of the surrogate loss.

lemma 2 *Suppose the perturbation amplitude* ρ *is sufficiently small, then the approximation to the surrogate gap in def 4.2 is always non-negative,* $h(\mathbf{w}) \approx f(\mathbf{w}^{adv}) - f(\mathbf{w}) \geq 0$

lemma 3 *For a local minima* $\mathbf{w}*$*, the surrogate gap is a measure of sharpness:* $\sigma_{max} \approx 2h(\mathbf{w}*)/\rho^2$.

Lemma 2 establishes that the surrogate gap is non-negative, while Lemma 3 shows that the loss landscape becomes flatter as $h \to 0$. Combining these, suggests that we can find regions with flat loss landscapes by minimising h.

The idea of GSAM is to simultaneously minimise the perturbed loss and the surrogate gap:

$$\min_{\mathbf{w}}(f_p(\mathbf{w}), h(\mathbf{w}))$$

By minimising $f_p(\mathbf{w})$ we seek a region with a low perturbed loss, like in SAM. By minimising $h(\mathbf{w})$ we search for a local minimum with a flat surface. A low perturbed loss implies low training loss within the neighbourhood and a flat loss surface reduces the generalisation gap [3]. Combining both of these we get high accuracy and good generalisation.

4.2 Gradient decompostion

Our goal is to minimise f_p and then to minimise h without increasing f_p. To do this, we decompose $\nabla(f(\mathbf{w})$ and ∇h into componenets that are parallel and orthogonal to $\nabla(f_p(\mathbf{w})$

$$\nabla f(\mathbf{w}_t) = \nabla f_{\|}(\mathbf{w}_t) + \nabla f_{\perp}(\mathbf{w}_t) \tag{1}$$

$$\nabla h(\mathbf{w}_t) = \nabla h_{\|}(\mathbf{w}_t) + \nabla h_{\perp}(\mathbf{w}_t) \tag{2}$$

$$\nabla h_{\perp}(\mathbf{w}_t) = -\nabla f_{\perp}(\mathbf{w}_t) \tag{3}$$

By construction, updating in the direction of $\nabla h_{\perp}(\mathbf{w}_t))$ does not change the value of f_p. This motivates the idea of performing a descent step in the $\nabla h_{\perp}(\mathbf{w}_t)$ direction. which is equivalent to an ascent step in the $\nabla f_{\perp}(\mathbf{w}_t)$ direction. This helps us achieve both of our goals simultaneously.

Algorithm 1 GSAM Algorithm

```
1: for $t = 1$ to $T$ do
2:     $\Delta w_t = \rho_t \frac{\nabla f^{(t)}}{\|\nabla f^{(t)}\|+\epsilon}$
3:     $w_t^{\text{adv}} = w_t + \Delta w_t$
4:     Obtain $\nabla f_p^{(t)}$ by backpropagation at $w_t^{\text{adv}}$.
5:     $\nabla f^{(t)} = \nabla f_{\|}^{(t)} + \nabla f_{\perp}^{(t)}$
6:     Update weights:
         • Vanilla: $w_{t+1} = w_t - \eta_t \nabla f^{(t)}$
         • SAM: $w_{t+1} = w_t - \eta_t \nabla f_p^{(t)}$
         • GSAM: $w_{t+1} = w_t - \eta_t \big(\nabla f_p^{(t)} - \alpha \nabla f_{\perp}^{(t)}\big)$
7: end for
```

Note that it is possible to adapt step 6 to involve momentum or adaptive learning rates as in [14].

It is tempting to try to minimise $\phi(\mathbf{w}) = \lambda f_p(\mathbf{w}) + (1-\lambda)h(\mathbf{w})$ However this does not lead to strong experimental results. This is because minimising f_p is far more important than minimising h, yet the two objectives are not necessarily aligned.

5 Adaptive Sharpness Aware Minimisation

In this section we explore an alternative approach to improving SAM by addressing its scale-dependency problem. We examine a solution, adaptive sharpness and use this concept to propose an improvement to GSAM.

Dinh et al. find that Sharpness defined in a rigid spherical region with a fixed radius can have a weak correlation with the generalization gap[5]. To rectify this we can take an adaptive approach to SAM, where the size of the neighbourhood ϵ adapts to ensure scale-invariance.

Motivated by these results, Kwon et al. note the following result[12]:

Theorem 2 *Let A be a scaling operator on the weight space that does not change the loss function.*

$$\max_{\|\epsilon\|\leq\rho} L_{\mathcal{S}}(\mathbf{w}+\epsilon) \neq \max_{\|\epsilon\|\leq\rho} L_{\mathcal{S}}(A\mathbf{w}+\epsilon)$$

Therefore, networks with parameters $\mathbf{w}$ and $A\mathbf{w}$ can have vastly different values of sharpness (as defined by the perturbed loss) yet have the same generalisation gap. We call this property scale dependency and it is one of the main causes for the weak correlation between sharpness and generalisation ability in SAM.

In order to explain the idea of adaptive sharpness, we must first define a normalisation operator.

Definition 5.1 *(Normalisation Operator) Let $\{T_{\mathbf{w}}, \mathbf{w} \in k\}$ be a family of invertible linear operators on $\mathbb{R}^k$. Given a weight $\mathbf{w}$, if $T^{-1}_{A\mathbf{w}}A = T^{-1}$ for any invertible scaling operator A on $\mathbb{R}^k$ which does not change the loss function, we say $T^{-1}_{\mathbf{w}}$ is a normalization operator of $\mathbf{w}$.*

Using the normalization operator, we define adaptive sharpness as follows:

Definition 5.2 *(Adaptive Sharpness) With $T_{\mathbf{w}}$ as a normalisation operator, the adaptive sharpness measure of $\mathbf{w}$ is defined as*

$$\max_{\|T_{\mathbf{w}}\epsilon\|_p\leq\rho} L_{\mathcal{S}}(\mathbf{w}+\epsilon) - L_{\mathcal{S}}(\mathbf{w})$$

for $1 \leq p \leq \infty$

We use as our normalisation operator $T_{\mathbf{w}} = diag\{w_1, w_2 \dots w_n\}$, We choose $p = 2$.

We can now define the ASAM algorithm. Similar to SAM, it follows a two step prodecure, where we now adjust the size of the neighbourhood using the normalisation operator to ensure scale-independence:

$$\begin{cases} \epsilon_t = \rho \frac{T_{\mathbf{w}}^2 \nabla L_{\mathcal{S}}(\mathbf{w}_t)}{\|T_{\mathbf{w}} \nabla L_{\mathcal{S}}(\mathbf{w}_t)\|_2} \\ \mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t(\nabla L_{\mathcal{S}}(\mathbf{w}_t + \epsilon_{\mathbf{t}}) + \lambda \mathbf{w}_t) \end{cases}$$

5.1 Adaptive-GSAM

In this section we describe a novel (to our knowledge) improvement on the GSAM algorithm which we call Adaptive-GSAM. We combine the surrogate loss measure from GSAM with the scale invariance of ASAM.

With $T_{\mathbf{w}}$ as defined in section 6, we adapt step 2 of the GSAM Algorithm to become:

$$\Delta \mathbf{w}_t = \rho_t \frac{\nabla f^{(t)}}{\|\nabla f^{(t)}\| + \epsilon} \cdot (|T_{\mathbf{w}}| + \gamma)$$

Where Gamma is a small hyper-parameter to prevent $\Delta \mathbf{w}_t$ values going to 0. By combining the scale invariance of ASAM with the improved sharpness metric used by GSAM, this algorithm achieves better generalisation results and reduces sensitivity to hyper-parameters.

6 Results

In this section we demonstrate Adaptive-GSAM's improved performance over vanilla GSAM. To implement Adaptive-GSAM, we modified a GSAM implementation built with TensorFlow [2], [1]. We first compare Adaptive GSAM to GSAM on the *moons* toy dataset in Python's scikit-learn module [13]. We use 10000 data points with a noise value of 0.3 and train for 50 epochs with a batch size of 64. We train on a 5 layer ReLU with 3 hidden layers of width 32.

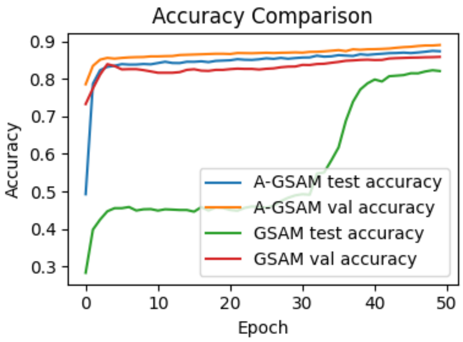

Figure 1: Moons dataset results

Method	Accuracy	Loss
Adaptive GSAM	89.0%	0.27
GSAM	85.7%	0.33

In this plot the test accuracy is higher than the train accuracy as the train accuracy is showing f_p, the worst case loss within the neighbourhood.
Early in training, GSAM achieves reasonable test performance but struggles with train accuracy, this suggests it is stuck in a region of high sharpness, whereas Adaptive-GSAM does not suffer from this problem.

We then compare performance on the MNIST [4] classification task using a simple CNN with 2 convolutional layers and 2 dense layers. We train for 25 epochs with a batch size of 32:

Method	Accuracy	Loss
Adaptive GSAM	98.5%	0.01
GSAM	95.3%	0.04

Figure 2: MNIST classification results

We see that Adaptive-GSAM outperforms GSAM. When selecting hyper-parameters, we noticed that Adaptive-GSAM demonstrated greater robustness to suboptimal choices. This could be attributed to its enhanced ability to adapt to the loss landscape.

7 Conclusion

GSAM and ASAM optimisation techniques yield superior generalisation performance compared to SAM. In this report we outlined the SAM, GSAM and ASAM algorithms, analysing their theoretical foundations and motivations. We then introduced the Adaptive-GSAM algorithm, showing it produces superior results to GSAM on two data classification tasks.

Our experiments were limited by computational resources restricting our tests to smaller datasets, future experiments could test Adaptive-GSAM on larger and more complex datasets such as CIFAR-10 or CIFAR-100. In addition the integration of SAM with momentum has shown promising results, suggesting that Adaptive-GSAM could also benefit from this approach.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] Marc Blanco. keras-gsam: Surrogate gap guided sharpness-aware minimization (gsam) implementation for keras/tensorflow 2. `https://github.com/mortfer/keras-gsam`, 2024. Accessed: 2024-12-27.

[3] 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. URL `https://arxiv.org/abs/1611.01838`.

[4] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[5] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets, 2017. URL `https://arxiv.org/abs/1703.04933`.

[6] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization, 2021. URL `https://arxiv.org/abs/2010.01412`.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. pages 529–536, 01 1994.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 01 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.1.1. URL `https://doi.org/10.1162/neco.1997.9.1.1`.

[9] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization, 2019. URL `https://arxiv.org/abs/1803.05407`.

[10] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them, 2019. URL `https://arxiv.org/abs/1912.02178`.

[11] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima, 2017. URL `https://arxiv.org/abs/1609.04836`.

[12] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks, 2021. URL `https://arxiv.org/abs/2102.11600`.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] Hao Sun, Li Shen, Qihuang Zhong, Liang Ding, Shixiang Chen, Jingwei Sun, Jing Li, Guangzhong Sun, and Dacheng Tao. Adasam: Boosting sharpness-aware minimization with adaptive learning rate and momentum for training deep neural networks, 2023. URL `https://arxiv.org/abs/2303.00565`.

[15] Ping yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=QC10RmRbZy9`.

[16] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training, 2022. URL `https://arxiv.org/abs/2203.08065`.