ADAPTIVE CURVATURE STEP SIZE: A PATH GEOMETRY BASED APPROACH TO OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose the Adaptive Curvature Step Size (ACSS) method, which dynamically adjusts the step size based on the local geometry of the optimization path. Our approach computes the normalized radius of curvature using consecutive gradients along the iterate path and sets the step-size equal to this radius. The effectiveness of ACSS stems from its ability to adapt to the local landscape of the optimization problem. In regions of low curvature, where consecutive gradient steps are nearly identical, ACSS allows for larger steps. Conversely, in areas of high curvature, where gradient steps differ significantly in direction, ACSS reduces the step size. This adaptive behavior enables more efficient navigation of complex loss landscapes. A key advantage of ACSS is its adaptive behavior based on local curvature information, which implicitly captures aspects of the function's second-order geometry without requiring additional memory. We provide a generalized framework for incorporating ACSS into various optimization algorithms, including SGD, Adam, AdaGrad, and RMSProp. Through extensive empirical evaluation on 20 diverse datasets, we compare ACSS variants against 12 popular optimization methods. Our results consistently show that ACSS provides performance benefits. Our results consistently show that ACSS provides performance benefits. We provide PyTorch implementations of ACSS versions for popular optimizers at our anonymized code repository.

1 INTRODUCTION

Optimization algorithms are the canonical work-horses of machine learning, driving the process of finding optimal parameters for deep learning models (Soydaner, 2020; Kochenderfer & Wheeler, 2019; Beck, 2017). As model architectures grow in size and complexity, the efficiency of these algorithms becomes paramount. A key challenge is that the objective in many learning problems are inherently non-convex, often due to structural or data-related constraints that impose non-convexity (Jain et al., 2017). Such learning problems may induce intricate loss landscapes characterized by large tracts of low gradients interspersed with areas of steep gradients, presenting significant navigational challenges for optimization algorithms. Effective optimization methods must not only find good solutions but do so efficiently in terms of computation and memory usage, especially when dealing with large-scale models and datasets, where navigation on the loss landscape is likely to follow an intricate path (Anil et al., 2019).

In light of this, we propose a geometric path based solution to optimization: the Adaptive Curvature Step Size (ACSS) method. Our approach is motivated by the observation that the curvature of the optimization path itself contains information about the local geometry of the loss landscape. By utilizing this curvature information, we can incorporate second order information adaptively into the step size — without the need for explicit computation or storage of second-order derivatives, and without the need for careful tuning of learning rates.

The intuition behind ACSS is rooted in differential geometry. Specifically, the curvature of a path provides insight into how rapidly the gradient is changing, which is indicative of the local shape of the loss surface. In fact, the iterate path can be viewed as a finite-difference approximation to the gradient flow manifold. We note that the curvature of this manifold is a powerful proxy for the local geometry of the loss landscape. Our method, ACSS, implicitly captures information about the changing gradient, which is related to the Hessian. This provides some of the benefits of second-order methods while maintaining the computational efficiency of first-order approaches.

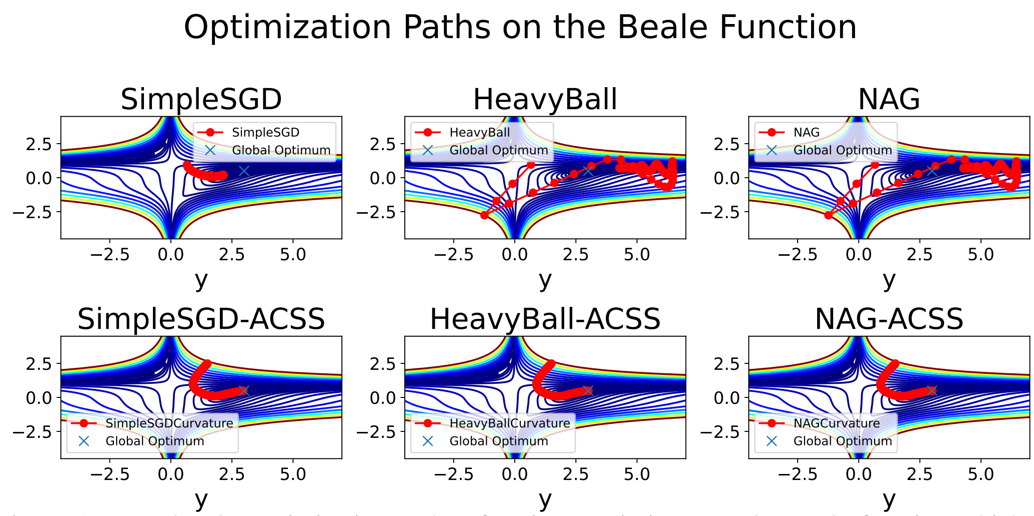


Figure 1: We plot the optimization paths of various optimizers on the Beale function which is characterized by steep valleys and a small area containing the global minimum. All optimizers start at $(-1.5, 2.5)$ with a learning rate of 1×10^{-3}. The function has a global minimum at $(3, 0.5)$; The ACSS versions of the optimizers converge here, without the use of any additional memory to store higher order moments.

1.1 Related Works:

First Order Methods: While first-order methods like Stochastic Gradient Descent (SGD) have low memory requirements, they converge slowly, particularly in ill-conditioned problems (Tian et al., 2023). Momentum based methods such as HeavyBall and NAG dampen oscillations to a certain degree (Sra et al., 2012; Nesterov, 2013), yet have limited ability to adapt when the loss landscape requires a change in direction of iterate (as seen in Figure 1).

Variance of Gradient: To address the limitations of basic SGD, several adaptive methods that adjust learning rates based on gradient statistics have been proposed. Adagrad accumulates squared gradients to adaptively tune learning rates, but it suffers from an ever-decreasing learning rate (Duchi et al., 2011). RMSProp improves upon this by using an exponentially decaying average of squared gradients, maintaining a more stable learning rate over time (Hinton et al., 2012). Adam and its variants (Kingma & Ba, 2014) further incorporate momentum, combining the benefits of adaptive learning rates and momentum to achieve better performance in various scenarios. AdamW enables better generalization through through weight decay regularization Loshchilov & Hutter (2017). AMSGrad addresses the convergence issues of Adam by ensuring that the learning rate does not increase, thereby providing better theoretical guarantees and more stable convergence in practice (Reddi et al., 2019). Nadam, and its weight decay variant NAdamW, integrate Nesterov momentum into the Adam framework, leading to faster convergence by anticipating the future position of

Optimizer	**Weights**	**Gradients**	**Momentum**	**Accumulated Squared Gradients**	**Exp. Avg. of Gradients**	**Exp. Avg. of Squared Gradients**
SimpleSGD	✓	✓	×	×	×	×
HeavyBall	✓	✓	✓	×	×	×
NAG	✓	✓	✓	×	×	×
Adagrad	✓	✓	×	✓	×	×
RMSProp	✓	✓	×	✓	×	×
Adadelta	✓	✓	×	×	×	✓
Adam	✓	✓	×	×	✓	✓
AdamW	✓	✓	×	×	✓	✓
AMSGrad	✓	✓	×	×	✓	✓
NAdam	✓	✓	×	×	✓	✓
NAdamW	✓	✓	×	×	✓	✓
RMSPropMomentum	✓	✓	✓	✓	×	×

Table 1: Memory requirements for different optimizers during backpropagation

the parameters (Dozat, 2016). However, these adaptive methods are not without drawbacks. They can sometimes lead to poor generalization (Wilson et al., 2017), and the implicit learning rate decay inherent in their designs can cause convergence issues in some scenarios (Reddi et al., 2019). Moreover, lack the ability to fully capture and utilize the local geometric information of the loss landscape, and often require careful tuning of hyper-parameters. We provide a study on the memory requirements of various optimizers in terms of the number of parameters in the model, in Table 1.

Second Order Methods: Second-order optimization methods typically offer better convergence properties, but Hessian based methods can get prohibitively expensive (Anil et al., 2020). Works like Gupta et al. (2018); Goldfarb et al. (2020); Singh et al. (2023) exploit the structure of the neural architecture that is being optimized (using factoring over layers) to reduce the computational cost, but these can face numerical instabilities. Subsequent works like Sophia (Liu et al., 2023) and AGD (Yue et al., 2023) address these issues, and yet have memory overhead. Recent works like Feinberg et al. (2024); Yen et al. (2024) address the memory issue to a certain degree, but they are essentially approximating the preconditioning tensor, which has a computation cost. Still other methods like VeLO (Metz et al., 2022) are frameworks that decide the optimization parameters using a small neural network — which has a wall-clock time overhead.

1.2 Our Contributions

1. **Novel Optimization Approach:** We introduce the Adaptive Curvature Step Size (ACSS) method, a new optimization algorithm that leverages the geometric properties of the optimization path to dynamically adjust step sizes. ACSS incorporates local curvature information derived from consecutive gradients, providing benefits typically associated with higher-order methods while maintaining the computational efficiency of first-order approaches. This approach allows ACSS to adapt to the local landscape of the optimization problem automatically, eliminating the need for careful manual tuning of step sizes typically required in traditional optimization methods.

2. **Low Memory Footprint with Performance Benefits:** Unlike many optimization methods that require significant additional memory for storing pre-conditioners or momentum terms, ACSS offers second-order benefits while maintaining the memory footprint of the base optimizer. Our experiments demonstrate that ACSS variants, particularly for optimizers like SGD, HeavyBall, and NAG that do not store squared gradients, show significant performance improvements across diverse datasets. For instance, SimpleSGD-ACSS often outperforms more complex methods like AdamW and AMSGrad, despite its lower memory requirements. This makes ACSS particularly suitable for large-scale optimization problems, where the reduced memory footprint can be leveraged to increase the number of parameters being optimized.

3. **Theoretical Foundation:** We provide a comprehensive theoretical analysis of ACSS, proving bounds on effective step size, stability under perturbations, convergence rates for strongly convex functions, and scale invariance properties. This analysis demonstrates ACSS's adaptive behavior to local curvature and offers insights into its relationship with both first-order and second-order optimization techniques.

4. **PyTorch Implementation:** To facilitate adoption and further research, we provide efficient PyTorch implementations of the ACSS variants for popular optimizers, at our anonymized GitHub repository, making it easy to incorporate our method into existing machine learning workflows and reproduce our results.

In the next section, we provide the necessary notations and theoretical machinery for ACSS.

2 Notations and Method

Consider a function $f : \mathbb{R}^n \times \mathcal{D} \to \mathbb{R}$ that we wish to minimize with respect to its first argument $w \in \mathbb{R}^n$. The optimization path traced by iterates $\{w_t\}$ can be viewed as a discrete approximation of a continuous curve in parameter space. Let $w_t \in \mathbb{R}^n$ be the parameter at iteration t, and $g_t = \nabla_w f(w_t, \mathcal{B}_t)$ be the gradient computed using a batch $\mathcal{B}_t \subset \mathcal{D}$.

In differential geometry, the curvature $\kappa(s)$ of a curve $w(s)$ parameterized by arc length s is defined as:

$$\kappa(s) = \left\| \frac{dT(s)}{ds} \right\|, \tag{1}$$

where $T(s) = \frac{dw(s)}{ds}$ is the unit tangent vector. The radius of curvature is given by $\rho(s) = \frac{1}{\kappa(s)}$.

To relate this to our discrete optimization steps, we approximate the curvature using finite differences. Let η be the base learning rate, and $g_t' = \nabla_w f(w_t - \eta g_t, \mathcal{B}_t)$ be the gradient at a *tentative* next point. We define the normalized radius of curvature as:

$$r_t := \frac{\|g_t\|}{\|g_t - g_t'\|}. \tag{2}$$

This approximation allows us to estimate the local curvature of the loss landscape without explicitly computing second-order derivatives.

To ensure numerical stability, we introduce a cap on the normalized radius of curvature:

$$\hat{r}_t := \min\{r_{\text{max}}, r_t\}, \tag{3}$$

where r_{max} is the maximum allowed curvature.

Update Rule: Incorporating this adaptive curvature step size, we define the update rule as:

$$w_{t+1} := w_t - \eta \times \hat{r}_t \times \frac{g_t}{\|g_t\|} \quad (\textbf{Eq. 1}) \tag{4}$$

This update can be interpreted as moving in the direction of the negative gradient $\frac{g_t}{\|g_t\|}$ with a step size dynamically adjusted by $\eta \times \hat{r}_t$ based on the local curvature of the loss landscape.

The proposed Adaptive Curvature Step Size (ACSS) method aims to balance the trade-off between convergence speed and stability by adapting the step size according to the geometry of the optimization path. In regions of low curvature, it allows for larger steps to accelerate progress, while in highly curved areas, it reduces the step size to maintain stability.

2.1 Algorithm

We now provide this update rule in the form of an Algorithm.

Algorithm 1: Stochastic gradient descent with adaptive curvature step size (SGD-ACSS)

Input: Function $f_w : \mathcal{D} \to \mathbb{R}$, initial parameters $w_0 \in \mathbb{R}^n$, base learning rate η, maximum radius r_{max}, number of iterations T, batch size B
Output: Optimized parameters w_T
for $t = 0$ **to** $T - 1$ **do**
  Sample a mini-batch $\mathcal{B}_t$ from $\mathcal{D}$;
  Compute gradient $g_t = \nabla f_w(w_t; \mathcal{B}_t)$;
  Compute tentative next point gradient $g_t' = \nabla f_w(w_t - \eta g_t; \mathcal{B}_t)$;
  Compute normalized radius of curvature $r_t = \frac{||g_t||}{||g_t - g_t'||}$;
  Compute capped radius $\hat{r}_t = \min\{r_{max}, r_t\}$;
  Update parameters $w_{t+1} = w_t - \eta \times \hat{r}_t \times \frac{g_t}{||g_t||}$;
end
return w_T

3 Theoretical Analysis

We provide theoretical guarantees for the Adaptive Curvature Step Size (ACSS) method. Our analysis focuses on the method's convergence properties, step size bounds, and adaptive behavior. Detailed proofs for all theorems can be found in the Appendix Section B.

3.1 Step Size Bounds and Convergence

We begin by establishing bounds on the effective step size of ACSS and proving its convergence for strongly convex functions.

Theorem 1 (Bounded Step Size of ACSS)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth and μ-strongly convex function. Consider the ACSS update rule with $r_{\max} \leq \frac{2}{\eta(\mu+L)}$. Then, the effective step size $\eta_{\text{eff}} = \eta \hat{r}_t$ is bounded as follows:

$$\frac{1}{L} \leq \eta_{\text{eff}} \leq \frac{2}{\mu + L}$$

for all iterations t.

This theorem ensures that ACSS maintains step sizes within a range that promotes stable convergence. Building on this result, we establish the convergence rate for ACSS:

Theorem 2 (Convergence Rate for ACSS on Strongly Convex Functions)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth and μ-strongly convex function. Under the ACSS update rule, for all $t \geq 0$:

$$\|w_t - w^*\|^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^t \|w_0 - w^*\|^2.$$

This theorem indicates that ACSS achieves linear convergence for strongly convex functions, with a rate comparable to standard gradient descent methods.

It is important to note that while the theoretical results presented in this section are derived for the deterministic gradient setting, the empirical results of ACSS, as discussed in Section 4, involves its use in stochastic settings with mini-batch optimization. The extension of these theoretical guarantees to the stochastic case is a potential area for future work. Nevertheless, our analysis does extend to scenarios involving bounded gradient perturbations, as detailed in the following subsection.

3.2 Stability under Perturbation

Next, we present results on the stability of ACSS under gradient perturbations and its convergence guarantees for L-smooth and μ-strongly convex functions.

Theorem 3 (Stability of ACSS Under Gradient Perturbations)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth and μ-strongly convex function. Assume the gradients are perturbed such that $\tilde{g}_t = g_t + \delta_t$ and $\tilde{g}'_t = g'_t + \delta'_t$, where $\|\delta_t\| \leq \varepsilon$ and $\|\delta'_t\| \leq \varepsilon$ for some $\varepsilon > 0$. Then, the difference between the updates using exact and perturbed gradients satisfies:

$$\|\tilde{w}_{t+1} - w_{t+1}\| \leq \frac{4\eta_{\max}\varepsilon}{m - \varepsilon},$$

where $\eta_{\max} = \frac{2}{L+\mu}$ and m is a lower bound on the gradient norm.

While this theoretical result provides partial insights under specific assumptions, it may not fully capture ACSS's behavior in complex, non-convex landscapes. However, our extensive experiments in Section 4 may provide further evidence of ACSS stability properties across several difficult-to-optimize problems and diverse common machine learning datasets.

3.3 Adaptive Behavior and Scale Invariance

Finally, we examine the scale invariance property of ACSS.

Theorem 4 (Scale Invariance of ACSS Effective Step Size)**.** For any scalar $\alpha > 0$, scaling the base step size η by α results in the same parameter updates for quadratic functions and approximately the same updates for general L-smooth and μ-strongly convex functions, assuming $r'_t \leq r_{\max}$.

This scale invariance property suggests that ACSS is not sensitive to the choice of base step size — a significant practical advantage. ACSS automatically adapts its effective step size to the local geometry of the loss landscape, taking larger steps in low-curvature regions and smaller steps in high-curvature areas. This behavior mitigates the need for manual step size tuning and allows ACSS to maintain near-optimal convergence rates across varying landscapes without requiring prior knowledge of function-specific parameters. In contrast, SGD often requires careful manual tuning of step sizes to achieve similar convergence rate guarantees, which is challenging, particularly when optimizing functions with varying curvature across the parameter space.

4 EXPERIMENTS

4.1 CROSS-DATASET PERFORMANCE ANALYSIS OF ACSS

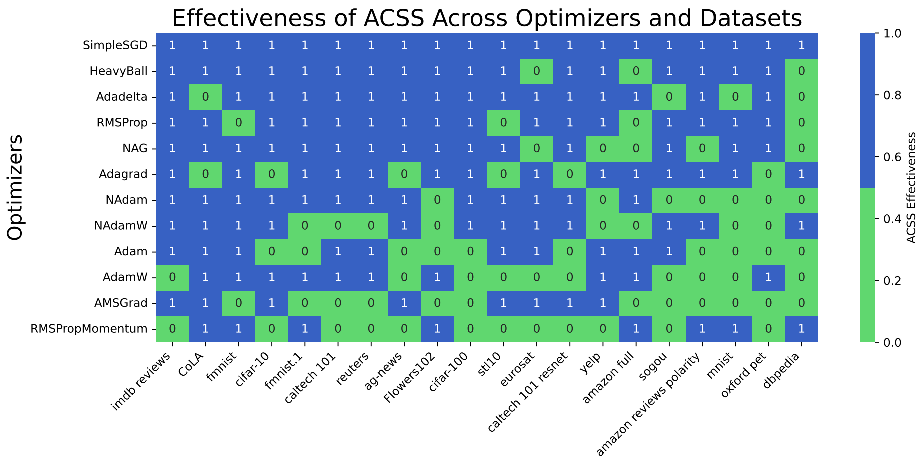


Figure 2: Binary representation of ACSS effectiveness across datasets and optimizers. Values indicate improvement (1) or no improvement (0) in training loss after a fixed number of epochs.

Figures 2 and 3 present a comprehensive evaluation of ACSS across 12 optimizers and 20 diverse datasets. ACSS demonstrates consistent performance improvements for most optimizer-dataset combinations. Significantly, SimpleSGD exhibits the most robust improvement across all datasets.

Optimizers that do not inherently use second-order information show the highest improvements, suggesting that ACSS effectively incorporates second-order information through loss landscape topology. SGD, HeavyBall, and NAG demonstrated mean training loss improvements of approximately 0.5 across 20 datasets using their respective ACSS versions.

Vision-related benchmarks, including Caltech 101, CIFAR-100, Flowers102, and STL10, showed the most significant improvements. The 18-layer ResNet variant exhibited the best performance, while the MNIST dataset with a simple neural network showed less pronounced improvements, likely due to the inherent effectiveness of most optimizers on simpler models.

Key Takeaways: ACSS provides improvements for most optimizers across various datasets. In cases where regular versions outperform ACSS, the difference in training loss is typically minimal.

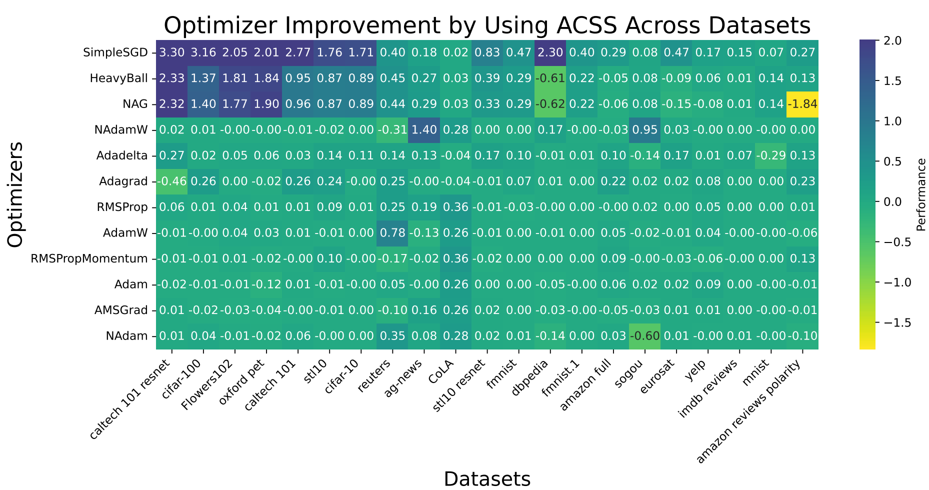


Figure 3: Quantitative improvement in training loss using ACSS across datasets and optimizers after a fixed number of epochs.

Table 2: Training Loss over 5 Epochs for Yelp Reviews Polarity Dataset (560,000 reviews) using a Simplified RNN Model. The model consists of embedding, RNN, and fully connected layers. ACSS versions of optimizers generally outperform their traditional counterparts.

Optimizer Name	Regular Optimizer					ACSS Version of Optimizer				
	Epoch 1	**Epoch 2**	**Epoch 3**	**Epoch 4**	**Epoch 5**	**Epoch 1**	**Epoch 2**	**Epoch 3**	**Epoch 4**	**Epoch 5**
Adadelta	0.680 ±0.00	0.674 ±0.00	0.671 ±0.00	0.669 ±0.00	0.668 ±0.00	0.679 ±0.01	0.670 ±0.00	0.666 ±0.00	0.663 ±0.00	0.659 ±0.00
Adagrad	0.558 ±0.01	0.521 ±0.01	0.510 ±0.01	0.501 ±0.01	0.493 ±0.01	0.569 ±0.07	0.498 ±0.07	0.452 ±0.06	0.429 ±0.06	0.410 ±0.07
Adam	0.627 ±0.01	0.584 ±0.01	0.587 ±0.00	0.568 ±0.04	0.575 ±0.02	0.542 ±0.04	0.541 ±0.16	0.530 ±0.17	0.457 ±0.20	0.489 ±0.14
AdamW	0.581 ±0.02	0.567 ±0.03	0.478 ±0.01	0.499 ±0.08	0.419 ±0.11	0.599 ±0.04	0.589 ±0.12	0.555 ±0.10	0.413 ±0.05	0.376 ±0.12
AMSGrad	0.537 ±0.00	0.548 ±0.01	0.569 ±0.11	0.481 ±0.05	0.589 ±0.07	0.616 ±0.04	0.596 ±0.02	0.625 ±0.08	0.625 ±0.03	0.578 ±0.03
HeavyBall	0.666 ±0.00	0.652 ±0.00	0.604 ±0.01	0.529 ±0.01	0.512 ±0.01	0.572 ±0.01	0.517 ±0.01	0.491 ±0.01	0.474 ±0.01	0.455 ±0.01
NAdam	0.637 ±0.01	0.612 ±0.00	0.589 ±0.00	0.580 ±0.04	0.537 ±0.09	0.609 ±0.02	0.543 ±0.05	0.543 ±0.01	0.531 ±0.04	0.538 ±0.02
NAdamW	0.601 ±0.01	0.531 ±0.00	0.495 ±0.05	0.498 ±0.05	0.523 ±0.03	0.632 ±0.00	0.594 ±0.02	0.585 ±0.02	0.541 ±0.04	0.528 ±0.02
NAG	0.666 ±0.00	0.652 ±0.00	0.604 ±0.01	0.529 ±0.01	0.510 ±0.02	0.630 ±0.02	0.616 ±0.00	0.604 ±0.01	0.604 ±0.03	0.591 ±0.02
RMSProp	0.650 ±0.04	0.538 ±0.07	0.495 ±0.13	0.425 ±0.09	0.447 ±0.03	0.624 ±0.02	0.493 ±0.03	0.432 ±0.03	0.407 ±0.06	0.394 ±0.06
RMSPropMomentum	0.652 ±0.02	0.578 ±0.04	0.561 ±0.03	0.491 ±0.05	0.467 ±0.03	0.633 ±0.06	0.601 ±0.03	0.581 ±0.00	0.551 ±0.04	0.524 ±0.04
SimpleSGD	0.676 ±0.00	0.671 ±0.00	0.669 ±0.00	0.667 ±0.00	0.665 ±0.00	0.596 ±0.01	0.535 ±0.01	0.519 ±0.01	0.506 ±0.01	0.493 ±0.02

4.2 Performance on the Yelp Reviews Dataset

We evaluated various optimizers with and without ACSS on the Yelp Reviews Polarity Dataset (560,000 reviews) using a simplified RNN model. The ACSS variants generally outperformed their standard counterparts over five epochs. AdamW-ACSS showed the most significant improvement, with loss decreasing from 0.5994 to 0.3756 across epochs, outperforming the traditional AdamW's final loss. SimpleSGD-ACSS demonstrated remarkable improvement, matching top performers like AdamW-ACSS by the first epoch.

Key Takeaways: The best performing non-ACSS optimizer after Epoch 5 reaches a training loss of only 0.419 (AdamW), which is reached at Epoch 4 for two of the ACSS versions. All the best-performing optimizers after Epoch 2 are ACSS versions of the optimizers.

4.3 Training Loss Improvements Averaged over all Datasets

We evaluated the performance of Adaptive Curvature Step Size (ACSS) variants of SimpleSGD, HeavyBall, and NAG (Nesterov Accelerated Gradient) across diverse datasets in vision and language domains. Our evaluation encompassed various model architectures, including CNNs (such as ResNet), RNNs, and simple neural networks. The results, as illustrated in Figure 4, demonstrate consistent improvements in training performance for ACSS variants compared to their standard counterparts. These improvements were observed across all five epochs and increased over time, indicating that ACSS provides sustained benefits throughout the training process.

Key Takeaways: Optimizers that do not store square-gradient terms (SGD, HeavyBall, NAG) exhibit significant outperformance through the use of ACSS. The improvement in mean training loss, averaged across all datasets, is evident across all the epochs.

4.4 Performance on vision benchmarks

Figure 5 presents a heatmap of optimizer rankings across five vision datasets: Caltech101, CIFAR10, Flowers102, MNIST, and STL10. The analysis reveals that Adadelta and RMSProp variants consistently underperform, with ACSS showing minimal impact on their effectiveness. In contrast, Adam, AdamW, and AMSGrad perform well initially, with ACSS offering marginal improvements. Adagrad demonstrates high performance variance across datasets.

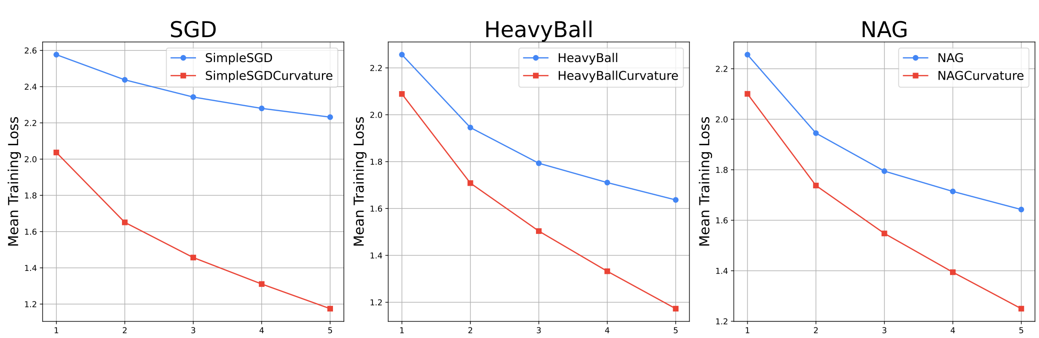


Figure 4: Mean training loss across epochs for different optimizers.

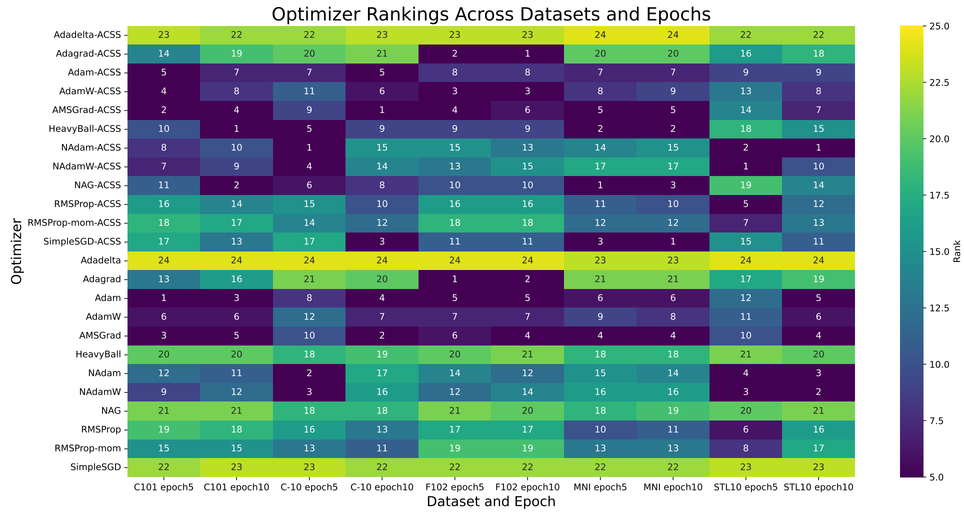


Figure 5: Heatmap of optimizer rankings across various computer vision datasets. The heatmap displays the performance ranks of 24 optimizers, including both standard versions and their Adaptive Curvature Step Size (ACSS) variants, on five different datasets (Caltech101, CIFAR10, Flowers102, MNIST, and STL10) at epochs 5 and 10. Rankings range from 1 (best performing) to 24 (worst performing), with lower numbers and cooler colors indicating better performance. This visualization highlights the impact of ACSS on various optimizers across different datasets.

Notably, optimizers that do not incorporate squared gradients (SimpleSGD, HeavyBall, NAG) benefit most from ACSS. These optimizers achieve performance boosts comparable to methods using squared gradients, but without the associated memory overhead.

Key Takeaways: ACSS versions generally outperform their traditional counterparts on these vision benchmarks for both ResNet-18 and simple CNN architectures. The most significant improvements are observed in optimizers that do not initially use squared gradients.

4.5 Overall rank improvements for different optimizers

Figure 6 illustrates the performance improvement of optimizers with ACSS across multiple datasets. Optimizers with lower memory requirements benefit most from ACSS. SimpleSGD, with the smallest memory footprint, shows the highest average rank improvement of 12.5. HeavyBall and NAG also demonstrate significant enhancements, with average improvements of 7.9 and 6.7 respectively.

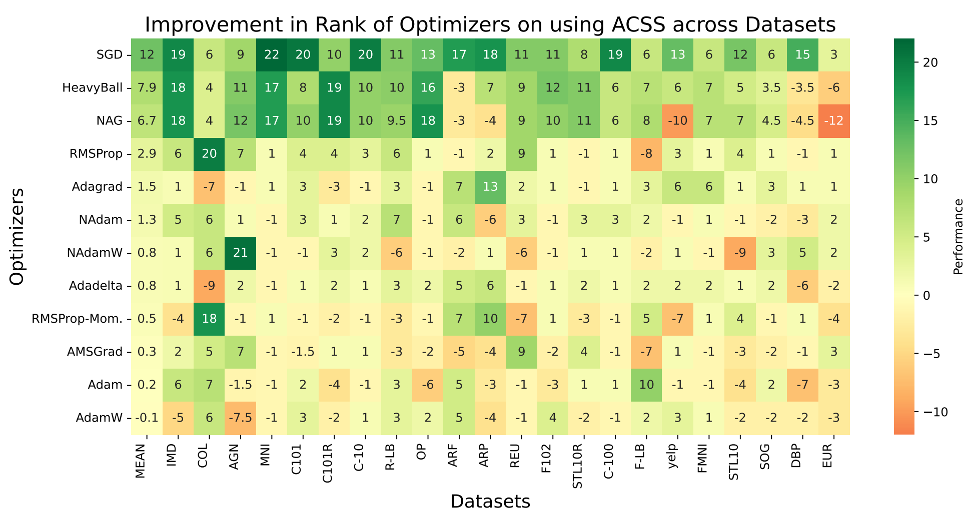
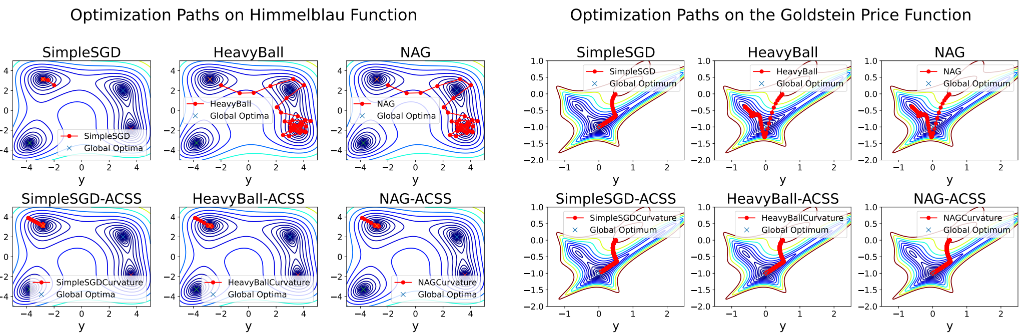


Figure 6: Heatmap of optimizer rank improvements when using ACSS across datasets. Green indicates better performance, red indicates worse. The datasets are listed on the X-axis, and the optimizers on the Y-axis. Color intensity represents the degree of improvement.



(a) (b)

Figure 7: Optimization paths on the Goldstein-Price (left) and Himmelblau (right) functions. These functions present challenges due to their complex landscapes with multiple optima and flat regions.

More complex optimizers like Adam, AdamW, and AMSGrad, which already incorporate adaptive learning rate mechanisms, show lower benefits. This suggests ACSS is particularly effective in enhancing simpler optimization algorithms, offering a memory-efficient alternative to more complex adaptive methods.

Key Takeaways: Except for AdamW, all optimizers show positive mean performance improvement with ACSS, indicating benefits in incorporating ACSS into existing optimization pipelines.

4.6 Optimization on challenging functions

We now plot the performance of our optimizers on two challenging functions: the Himmelblau and Goldstein-Price functions. Additional functions are analyzed in Appendix F.

The Himmelblau Function: The Himmelblau function has four global minima. ACSS versions converge to the nearest minimum from the starting point (-4,4), while other versions overshoot at a learning rate of 1.5×10^{-2}. At higher rates, non-ACSS versions diverge, whereas ACSS versions maintain convergence.

The Goldstein-Price Function: The Goldstein-Price function, with its complex landscape of multiple local minima and one global minimum at (0, -1), challenges gradient-based methods. ACSS optimizers dynamically adjust step sizes based on local curvature, enabling precise convergence to the global minimum. In contrast, standard Heavyball and NAG optimizers overshoot, moving toward different local minima. We plot 5000 iterations from (0.5, 0) with a learning rate of 2.5×10^{-5}.

Key Takeaways: In Figures 1, 7 in the main paper, and Figure 8 in Appendix F, we plot the ACSS performance as compared with the regular versions for challenging optimization benchmark functions. In all the cases, the ACSS versions showed better stability and convergence properties compared to the traditional algorithms.

4.7 Limitations:

It is important to acknowledge that ACSS introduces additional computational overhead per iteration, with theoretical analysis suggesting up to twice the cost and experimental wall-clock time measurements showing an average increase of 1.37 times for the ACSS optimizers over their traditional counterparts, which is balanced against its memory efficiency benefits and lower time to convergence (see Section D for detailed theoretical and experimental analyses).

5 Conclusions

This work introduced the Adaptive Curvature Step Size (ACSS) method, a novel optimization approach that leverages the geometric properties of the optimization path to dynamically adjust step sizes. Our comprehensive empirical evaluation across diverse datasets and challenging functions demonstrates that ACSS consistently outperforms traditional optimization methods. The method's ability to incorporate second-order-like information without explicit computation of the Hessian is a key benefit, as we show through our theoretical guarantees. Furthermore, ACSS's low memory footprint makes it particularly suitable for large-scale optimization setups and low-resource settings. The generalized framework we provide for incorporating ACSS into various optimization algorithms, along with our PyTorch implementations, facilitates further research in this direction.

REFERENCES

Rohan Anil, Vineet Gupta, Tomer Koren, and Yoram Singer. Memory efficient adaptive optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable second order optimization for deep learning. *arXiv preprint arXiv:2002.09018*, 2020.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Timothy Dozat. Incorporating nesterov momentum into adam. *Stanford CS 229 Project*, 2016.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Vladimir Feinberg, Xinyi Chen, Y Jennifer Sun, Rohan Anil, and Elad Hazan. Sketchy: Memory-efficient adaptive regularization with frequent directions. *Advances in Neural Information Processing Systems*, 36, 2024.

Donald Goldfarb, Yi Ren, and Achraf Bahamou. Practical quasi-newton methods for training deep neural networks. *Advances in Neural Information Processing Systems*, 33:2386–2396, 2020.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

Prateek Jain, Purushottam Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–363, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Mykel J Kochenderfer and Tim A Wheeler. *Algorithms for optimization*. Mit Press, 2019.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *arXiv preprint arXiv:2305.14342*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Luke Metz, James Harrison, C Daniel Freeman, Amil Merchant, Lucas Beyer, James Bradbury, Naman Agrawal, Ben Poole, Igor Mordatch, Adam Roberts, et al. Velo: Training versatile learned optimizers by scaling up. *arXiv preprint arXiv:2211.09760*, 2022.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Siddharth Singh, Zachary Sating, and Abhinav Bhatele. Jorge: Approximate preconditioning for gpu-efficient second-order optimization. *arXiv preprint arXiv:2310.12298*, 2023.

Derya Soydaner. A comparison of optimization algorithms for deep learning. *International Journal of Pattern Recognition and Artificial Intelligence*, 34(13):2052013, 2020.

Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. *Optimization for machine learning*. Mit Press, 2012.

Yingjie Tian, Yuqi Zhang, and Haibin Zhang. Recent advances in stochastic gradient descent in deep learning. *Mathematics*, 11(3):682, 2023.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017.

Jui-Nan Yen, Sai Surya Duvvuri, Inderjit Dhillon, and Cho-Jui Hsieh. Block low-rank preconditioner with shared basis for stochastic optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

Yun Yue, Zhiling Ye, Jiadi Jiang, Yongchao Liu, and Ke Zhang. Agd: an auto-switchable optimizer using stepwise gradient difference for preconditioning matrix. *Advances in Neural Information Processing Systems*, 36:45812–45832, 2023.

SUPPLEMENTARY MATERIALS

These supplementary materials provide additional details, derivations, and experimental results for our paper. The appendix is organized as follows:

- Section A presents detailed derivations of the Adaptive Curvature Step Size (ACSS) method.
- Section B offers a comprehensive theoretical analysis of ACSS, including proofs of key theorems.
- Section C introduces a generalized algorithm for incorporating ACSS into existing optimizers.
- Section D provides theoretical and experimental analyses pertaining to limitations of this work.
- Section E provides additional experimental results, like performance on the CoLA dataset.
- Section F details the testing functions used to benchmark ACSS optimization.

A DETAILED DERIVATIONS OF ACSS

The optimization path traced by iterates $\{w_t\}$ during the optimization process can be viewed as a discrete approximation of a continuous curve in parameter space. Understanding the curvature of this path provides valuable insights into the local geometry of the loss landscape and guides adaptive step size selection. In differential geometry, the curvature $\kappa(s)$ of a curve $w(s)$ parameterized by arc length s is defined as:

$$\kappa(s) = \left\| \frac{dT(s)}{ds} \right\|, \tag{5}$$

where $T(s) = \frac{dw(s)}{ds}$ is the unit tangent vector to the curve at point s. The radius of curvature $\rho(s)$ is then given by $\rho(s) = \frac{1}{\kappa(s)}$.

In the context of gradient-based optimization, we consider the continuous-time dynamics governed by the gradient flow:

$$\frac{dw(t)}{dt} = -\nabla f(w(t)) = -g(t), \tag{6}$$

where $g(t) = \nabla f(w(t))$ is the gradient of the function f at $w(t)$. To relate curvature to discrete optimization steps, we approximate the curvature using finite differences. We define the unit tangent vector at iteration t as $T_t = -\frac{g_t}{\|g_t\|}$, and approximate the change in the unit tangent vector between iterations t and $t+1$ as $\Delta T_t \approx -\frac{g_{t+1} - g_t}{\|g_t\|}$.

The curvature κ_t at iteration t, given this gradient norm approximation can then be given as:

$$\kappa_t = \frac{\|g_{t+1} - g_t\|}{\|g_t\|\eta}, \tag{7}$$

where η is the step size. Consequently, the radius of curvature ρ_t is:

$$\rho_t = \frac{1}{\kappa_t} = \frac{\|g_t\|\eta}{\|g_{t+1} - g_t\|}. \tag{8}$$

We introduce a normalized radius of curvature $r_t = \frac{\rho_t}{\eta} = \frac{\|g_t\|}{\|g_t' - g_t\|}$, which decouples the radius of curvature from the base learning rate η. The adaptive step size Δs_t is then defined as $\Delta s_t = \eta \times r_t = \eta \times \frac{\|g_t\|}{\|g_t' - g_t\|}$. To maintain numerical stability, we introduce a cap on the normalized radius of curvature: $\hat{r}_t = \min\{r_{\text{max}}, r_t\}$, where r_{max} is a predefined maximum radius of curvature.

A.1 Final Update Rule and Discussion

The final parameter update rule for the Adaptive Curvature Step Size (ACSS) method is:

$$w_{t+1} = w_t - \eta \times \hat{r}_t \times \frac{g_t}{\|g_t\|}. \tag{9}$$

This can be interpreted as moving in the direction of the negative gradient $\frac{g_t}{\|g_t\|}$ with a step size scaled by $\eta \times \hat{r}_t$.

The ACSS method offers several key advantages in optimization tasks. By leveraging the curvature of the optimization path, it implicitly incorporates second-order information without the computational overhead of explicit second-order methods. This dynamic adaptation allows ACSS to navigate complex loss landscapes more effectively, enabling rapid progress in flat regions while ensuring stability in high-curvature areas. The method's memory efficiency, requiring minimal additional storage beyond current and tentative gradients, makes it particularly suitable for large-scale optimization problems in deep learning. Furthermore, ACSS's framework allows for integration into various existing optimization algorithms such as SGD, Adam, AdaGrad, and RMSProp, enhancing their performance with its curvature-based step size adjustment.

B Theoretical Analysis

Theorem 5 (Bounded Step Size of ACSS)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth and μ-strongly convex function. Consider the ACSS update rule $w_{t+1} = w_t - \eta\hat{r}_t \frac{g_t}{\|g_t\|}$ where $\hat{r}_t = \min\{r_{\max}, r_t\}$ and $r_t = \frac{\|g_t\|}{\|g_t - g_t'\|}$. Assume the following:

1. The gradients are bounded: $\exists G > 0$ such that $\|g_t\| \leq G$ for all t
2. The function f is L-smooth: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all x, y
3. The function f is μ-strongly convex: $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2$ for all x, y
4. The maximum radius $r_{\max}$ is chosen such that $r_{\max} \leq \frac{2}{\eta(\mu+L)}$

Then, the effective step size $\eta_{\text{eff}} = \eta\hat{r}_t$ is bounded as follows:

$$\frac{1}{L} \leq \eta_{\text{eff}} \leq \frac{2}{\mu + L}$$

for all iterations t.

Proof. We proceed as follows:

$$r_t = \frac{\|g_t\|}{\|g_t - g_t'\|} \qquad \text{Definition of } r_t$$

$$g_t' = \nabla f(w_t - \eta g_t) \qquad \text{From the algorithm}$$

$$\begin{aligned}\|g_t - g_t'\| &= \|\nabla f(w_t) - \nabla f(w_t - \eta g_t)\| \\ &\leq L\|\eta g_t\| = L\eta\|g_t\| \qquad \text{Using } L\text{-smoothness}\end{aligned}$$

$$r_t = \frac{\|g_t\|}{\|g_t - g_t'\|} \geq \frac{\|g_t\|}{L\eta\|g_t\|} = \frac{1}{L\eta} \qquad \text{Lower bound on } r_t$$

$$\hat{r}_t = \min\{r_{\max}, r_t\} \geq \min\{\frac{2}{\eta(\mu+L)}, \frac{1}{L\eta}\} = \frac{1}{L\eta} \qquad \text{Since } \frac{2}{\mu+L} > \frac{1}{L}$$

$$\eta_{\text{eff}} = \eta\hat{r}_t \geq \eta\frac{1}{L\eta} = \frac{1}{L} \qquad \text{Lower bound on } \eta_{\text{eff}}$$

$$\eta_{\text{eff}} = \eta\hat{r}_t \leq \eta r_{\max} \leq \eta\frac{2}{\eta(\mu+L)} = \frac{2}{\mu+L} \qquad \text{Upper bound on } \eta_{\text{eff}}$$

Thus, we have established that $\frac{1}{L} \leq \eta_{\text{eff}} \leq \frac{2}{\mu+L}$ for all iterations t. □

Theorem 6 (Convergence of Gradient Descent on Quadratic Functions)**.** Consider the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ defined as

$$f(w) = \frac{1}{2}w^T Aw - b^T w + c,$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite with eigenvalues $0 < \mu \leq \lambda_1 \leq \cdots \leq \lambda_n \leq L$, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. For the gradient descent update rule with step size $\eta_{\text{eff}} > 0$:

$$w_{t+1} = w_t - \eta_{\text{eff},t}\nabla f(w_t) = w_t - \eta_{\text{eff},t}(Aw_t - b),$$

convergence is guaranteed if and only if $0 < \eta_{\text{eff}} < \frac{2}{\lambda_n}$. Moreover, the optimal convergence rate is achieved when $\eta_{\text{eff}} = \frac{2}{\mu+L}$.

Proof. The gradient of f is $\nabla f(w) = Aw - b$, yielding the unique minimizer $w^* = A^{-1}b$. Let $e_t = w_t - w^*$ denote the error at step t. The update rule can be rewritten as:

$$e_{t+1} = (I - \eta_{\text{eff},t}A)e_t$$

Since A is symmetric positive definite, it can be diagonalized as $A = Q\Lambda Q^T$, where Q is orthogonal and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Define $\tilde{e}_t = Q^T e_t$. Then:

$$\tilde{e}_{t+1} = (I - \eta_{\text{eff},t}\Lambda)\tilde{e}_t$$

This implies that for each component i:

$$\tilde{e}^i_{t+1} = (1 - \eta_{\text{eff},t}\lambda_i)\tilde{e}^i_t$$

For convergence, we require $|1 - \eta_{\text{eff}}\lambda_i| < 1$ for all i, which leads to:

$$0 < \eta_{\text{eff}} < \frac{2}{\lambda_i} \quad \forall i$$

Since λ_n is the largest eigenvalue, the condition $0 < \eta_{\text{eff}} < \frac{2}{\lambda_n}$ ensures convergence.

The convergence rate is determined by $\max_i |1 - \eta_{\text{eff}}\lambda_i|$. To minimize this, we solve:

$$\min_{\eta_{\text{eff}}} \max\{|1 - \eta_{\text{eff}}\mu|, |1 - \eta_{\text{eff}}L|\}$$

The optimal solution occurs when $1 - \eta_{\text{eff}}\mu = -(1 - \eta_{\text{eff}}L)$, yielding $\eta_{\text{eff}} = \frac{2}{\mu+L}$.

Therefore, gradient descent converges if and only if $0 < \eta_{\text{eff}} < \frac{2}{\lambda_n}$, with the optimal convergence rate achieved at $\eta_{\text{eff}} = \frac{2}{\mu+L}$. □

Theorem 7 (Convergence of Gradient Descent on L-Smooth and μ-Strongly Convex Functions)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth and μ-strongly convex function. For the gradient descent update rule with step size $\eta_{\text{eff}} > 0$:

$$w_{t+1} = w_t - \eta_{\text{eff},t}\nabla f(w_t),$$

convergence to the unique minimizer w^* is optimally achieved when $\eta_{\text{eff}} = \frac{2}{\mu+L}$.

Proof. Given that f is L-smooth and μ-strongly convex, we have:

$$\mu I \preceq \nabla^2 f(w) \preceq LI \quad \forall w \in \mathbb{R}^n.$$

Let w^* be the unique minimizer of f. Define the error vector $e_t = w_t - w^*$. The gradient descent update can be written as:

$$e_{t+1} = e_t - \eta_{\text{eff},t}\nabla f(w_t).$$

By the Mean Value Theorem, there exists ξ_t on the line segment between w_t and w^* such that:

$$\nabla f(w_t) = \nabla^2 f(\xi_t)e_t.$$

Thus, we can rewrite the error dynamics as:

$$e_{t+1} = (I - \eta_{\text{eff},t}\nabla^2 f(\xi_t))e_t.$$

Taking the Euclidean norm and using the operator norm:

$$\|e_{t+1}\| \leq \|I - \eta_{\text{eff},t}\nabla^2 f(\xi_t)\| \cdot \|e_t\|.$$

The eigenvalues of $\nabla^2 f(\xi_t)$ lie in $[\mu, L]$ by Lemma 1. For convergence, we require:

$$|1 - \eta_{\text{eff}}\lambda| < 1 \quad \forall \lambda \in [\mu, L].$$

Similar to Theorem 6, the convergence rate is determined by $\max_{\lambda\in[\mu,L]} |1 - \eta_{\text{eff}}\lambda|$. To minimize this, we solve:

$$\min_{\eta_{\text{eff}}} \max\{|1 - \eta_{\text{eff}}\mu|, |1 - \eta_{\text{eff}}L|\}$$

The optimal solution occurs when $1 - \eta_{\text{eff}}\mu = -(1 - \eta_{\text{eff}}L)$, yielding $\eta_{\text{eff}} = \frac{2}{\mu+L}$.

Therefore, gradient descent converges if $0 < \eta_{\text{eff}} < \frac{2}{\mu+L}$. □

Lemma 1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth and μ-strongly convex function. Then for any $\xi \in \mathbb{R}^n$, the eigenvalues of the Hessian matrix $\nabla^2 f(\xi)$ lie in the interval $[\mu, L]$.

Proof. We begin by establishing that $\mu I \preceq \nabla^2 f(\xi) \preceq LI$ for all $\xi \in \mathbb{R}^n$, where $\preceq$ denotes the semidefinite ordering and I is the identity matrix. From this, we will conclude that the eigenvalues of $\nabla^2 f(\xi)$ lie in $[\mu, L]$.

First, consider the L-smoothness property. For any $x, y \in \mathbb{R}^n$, we have:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

For an arbitrary direction $v \in \mathbb{R}^n$, this implies:

$$\lim_{t \to 0} \frac{\|\nabla f(x + tv) - \nabla f(x)\|}{t} \le L\|v\|$$

Taking the limit, we obtain:

$$\|[\nabla^2 f(x)]v\| \le L\|v\|$$

This inequality is equivalent to:

$$v^T[\nabla^2 f(x)]v \le Lv^T v \quad \forall v \in \mathbb{R}^n$$

which can be expressed in matrix notation as $\nabla^2 f(x) \preceq LI$.

Now, we turn to the μ-strong convexity property. For any $x, y \in \mathbb{R}^n$:

$$(\nabla f(x) - \nabla f(y))^T (x - y) \ge \mu\|x - y\|^2$$

Following a similar argument as above, we can show that:

$$v^T[\nabla^2 f(x)]v \ge \mu v^T v \quad \forall v \in \mathbb{R}^n$$

which is equivalent to $\nabla^2 f(x) \succeq \mu I$.

Combining these results, we have established that for all $\xi \in \mathbb{R}^n$:

$$\mu I \preceq \nabla^2 f(\xi) \preceq LI$$

Now, we invoke a fundamental result from linear algebra: for any symmetric matrix A, the statement $\lambda I \preceq A \preceq \Lambda I$ is equivalent to $\lambda \le \lambda_i(A) \le \Lambda$ for all eigenvalues $\lambda_i(A)$ of A. Since $\nabla^2 f(\xi)$ is symmetric (due to the assumed twice differentiability of f), we can apply this result.

Therefore, we conclude that for any $\xi \in \mathbb{R}^n$, all eigenvalues λ_i of $\nabla^2 f(\xi)$ satisfy:

$$\mu \le \lambda_i \le L$$

Thus, the eigenvalues of $\nabla^2 f(\xi)$ lie in the interval $[\mu, L]$, completing the proof. □

Theorem 8 (Stability of ACSS Under Gradient Perturbations)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth and μ-strongly convex function. Consider the ACSS update rule:

$$w_{t+1} = w_t - \eta_{\text{eff},t} \frac{g_t}{\|g_t\|},$$

where:

$$\eta_{\text{eff},t} = \eta \hat{r}_t, \quad \hat{r}_t = \min\{r_{\max}, r_t\}, \quad r_t = \frac{\|g_t\|}{\|g_t - g_t'\|}, \quad g_t' = \nabla f(w_t - \eta g_t).$$

Assume the following:

1. The gradients are bounded: $\exists G > m > 0$ such that $m \le \|g_t\| \le G$ for all t.
2. The function f is L-smooth: $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$ for all $x, y \in \mathbb{R}^n$.
3. The function f is μ-strongly convex: $\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \mu\|x - y\|^2$ for all $x, y \in \mathbb{R}^n$.
4. The maximum radius $r_{\max}$ is chosen such that $r_{\max} \le \frac{2}{(L+\mu)\eta}$.
5. Gradients are perturbed: $\tilde{g}_t = g_t + \delta_t$ and $\tilde{g}_t' = g_t' + \delta_t'$, where $\|\delta_t\| \le \varepsilon$ and $\|\delta_t'\| \le \varepsilon$ for some $\varepsilon > 0$.

Then, the difference between the updates using exact and perturbed gradients satisfies:

$$\|w_{t+1} - \tilde{w}_{t+1}\| \le \frac{4\eta_{\max}\varepsilon}{m - \varepsilon},$$

where $\eta_{\max} = \frac{2}{L+\mu}$.

Proof. We begin by expressing the difference between the exact update w_{t+1} and the perturbed update $\tilde{w}_{t+1}$:

$$\|w_{t+1} - \tilde{w}_{t+1}\| = \left\| \eta_{\text{eff},t} \frac{g_t}{\|g_t\|} - \tilde{\eta}_{\text{eff},t} \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\|$$

Applying Lemma 2, we obtain:

$$\|w_{t+1} - \tilde{w}_{t+1}\| \le \eta_{\text{eff},t} \left\| \frac{g_t}{\|g_t\|} - \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\| + |\eta_{\text{eff},t} - \tilde{\eta}_{\text{eff},t}|$$

We now bound each term separately.

First, we bound the difference in direction vectors. Note that $\tilde{g}_t = g_t + \delta_t$, so $\|g_t - \tilde{g}_t\| = \|\delta_t\| \le \varepsilon$. Also, from our assumptions, $\|g_t\| \ge m > \varepsilon$. Therefore, we can apply Proposition B.1 with $a = g_t$ and $b = \tilde{g}_t$:

$$\left\| \frac{g_t}{\|g_t\|} - \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\| \le \frac{2\|g_t - \tilde{g}_t\|}{\|g_t\|} \le \frac{2\varepsilon}{m}$$

Next, we bound the difference in effective step sizes:

$$|\eta_{\text{eff},t} - \tilde{\eta}_{\text{eff},t}| = \eta|\hat{r}_t - \tilde{\hat{r}}_t| \le \eta|r_t - \tilde{r}_t|$$

To bound $|r_t - \tilde{r}_t|$, we use the definition of r_t and $\tilde{r}_t$:

$$|r_t - \tilde{r}_t| = \left| \frac{\|g_t\|}{\|g_t - g_t'\|} - \frac{\|\tilde{g}_t\|}{\|\tilde{g}_t - \tilde{g}_t'\|} \right|$$

Using the triangle inequality and the fact that $\|\tilde{g}_t\| \le \|g_t\| + \|\delta_t\| \le G + \varepsilon$, and $\|\tilde{g}_t - \tilde{g}_t'\| \ge \|g_t - g_t'\| - 2\varepsilon$, we can derive:

$$|r_t - \tilde{r}_t| \le \frac{2\varepsilon}{m - 2\varepsilon}$$

Thus,

$$|\eta_{\text{eff},t} - \tilde{\eta}_{\text{eff},t}| \le \frac{2\eta\varepsilon}{m - 2\varepsilon}$$

Combining these bounds and using $\eta_{\text{eff},t} \le \eta_{\max} = \frac{2}{L+\mu}$, we arrive at:

$$\|w_{t+1} - \tilde{w}_{t+1}\| \le \eta_{\max} \cdot \frac{2\varepsilon}{m} + \frac{2\eta\varepsilon}{m - 2\varepsilon}$$

Given the choice of $r_{\max}$ and the boundedness of $\eta_{\text{eff},t}$, we can bound this as:

$$\|w_{t+1} - \tilde{w}_{t+1}\| \le \frac{2(\eta_{\max} + \eta)\varepsilon}{m - \varepsilon} \le \frac{4\eta_{\max}\varepsilon}{m - \varepsilon}$$

In practice, η is far lower than $\eta_{\max}$, and hence we can ignore the second term in which case we get a tighter bound of

$$\|w_{t+1} - \tilde{w}_{t+1}\| \leq \frac{2\eta_{\max}\varepsilon}{m}$$

This bound demonstrates that the ACSS algorithm is stable under bounded gradient perturbations, with the perturbation in the parameter updates being proportional to the noise level ε and inversely proportional to $m - \varepsilon$. ☐

Lemma 2 (Triangle Inequality for ACSS Updates)**.** Given the ACSS update rule and its perturbed version:

$$w_{t+1} = w_t - \eta_{\text{eff},t} \frac{g_t}{\|g_t\|}, \quad \tilde{w}_{t+1} = w_t - \tilde{\eta}_{\text{eff},t} \frac{\tilde{g}_t}{\|\tilde{g}_t\|}$$

The difference between these updates can be bounded as:

$$\|w_{t+1} - \tilde{w}_{t+1}\| \leq \eta_{\text{eff},t} \left\| \frac{g_t}{\|g_t\|} - \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\| + |\eta_{\text{eff},t} - \tilde{\eta}_{\text{eff},t}|$$

Proof. We start with the difference between the updates:

$$\|w_{t+1} - \tilde{w}_{t+1}\| = \left\| \eta_{\text{eff},t} \frac{g_t}{\|g_t\|} - \tilde{\eta}_{\text{eff},t} \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\|$$

Add and subtract $\eta_{\text{eff},t} \frac{\tilde{g}_t}{\|\tilde{g}_t\|}$ inside the norm:

$$\|w_{t+1} - \tilde{w}_{t+1}\| = \left\| \eta_{\text{eff},t} \frac{g_t}{\|g_t\|} - \eta_{\text{eff},t} \frac{\tilde{g}_t}{\|\tilde{g}_t\|} + \eta_{\text{eff},t} \frac{\tilde{g}_t}{\|\tilde{g}_t\|} - \tilde{\eta}_{\text{eff},t} \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\|$$

Apply the triangle inequality:

$$\|w_{t+1} - \tilde{w}_{t+1}\| \leq \left\| \eta_{\text{eff},t} \frac{g_t}{\|g_t\|} - \eta_{\text{eff},t} \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\| + \left\| \eta_{\text{eff},t} \frac{\tilde{g}_t}{\|\tilde{g}_t\|} - \tilde{\eta}_{\text{eff},t} \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\|$$

Factor out $\eta_{\text{eff},t}$ from the first term and simplify the second term:

$$\|w_{t+1} - \tilde{w}_{t+1}\| \leq \eta_{\text{eff},t} \left\| \frac{g_t}{\|g_t\|} - \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\| + \left\| (\eta_{\text{eff},t} - \tilde{\eta}_{\text{eff},t}) \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\|$$

Note that $\left\| \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\| = 1$, so:

$$\|w_{t+1} - \tilde{w}_{t+1}\| \leq \eta_{\text{eff},t} \left\| \frac{g_t}{\|g_t\|} - \frac{\tilde{g}_t}{\|\tilde{g}_t\|} \right\| + |\eta_{\text{eff},t} - \tilde{\eta}_{\text{eff},t}|$$

This completes the proof of the lemma. ☐

Proposition B.1 (Bound on Difference of Normalized Vectors)**.** Given two vectors $a, b \in \mathbb{R}^n$ with $\|a\| > \|a - b\|$, we have:

$$\left\| \frac{a}{\|a\|} - \frac{b}{\|b\|} \right\| \leq \frac{2\|a - b\|}{\|a\|}$$

Proof. We start with the vector identity:

$$\frac{a}{\|a\|} - \frac{b}{\|b\|} = \frac{a\|b\| - b\|a\|}{\|a\|\|b\|} = \frac{a(\|b\| - \|a\|) + \|a\|(a - b)}{\|a\|\|b\|}$$

Taking the norm of both sides and applying the triangle inequality:

$$\left\| \frac{a}{\|a\|} - \frac{b}{\|b\|} \right\| \leq \frac{\|a\|\|\|b\| - \|a\|\| + \|a\|\|a - b\|}{\|a\|\|b\|}$$

Using the reverse triangle inequality, $|\|b\| - \|a\|| \leq \|b-a\| = \|a-b\|$:

$$\left\| \frac{a}{\|a\|} - \frac{b}{\|b\|} \right\| \leq \frac{\|a\|\|a-b\| + \|a\|\|a-b\|}{\|a\|\|b\|} = \frac{2\|a-b\|}{\|b\|}$$

Since $\|b\| \geq \|a\| - \|a-b\|$ (by the triangle inequality), and given $\|a\| > \|a-b\|$, we have:

$$\left\| \frac{a}{\|a\|} - \frac{b}{\|b\|} \right\| \leq \frac{2\|a-b\|}{\|a\| - \|a-b\|} \leq \frac{2\|a-b\|}{\|a\|}$$

This completes the proof of the lemma. $\square$

Theorem 9 (Convergence Rate for ACSS on Strongly Convex Functions)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth and μ-strongly convex function. Consider the ACSS update rule:

$$w_{t+1} = w_t - \eta_{\text{eff},t} \frac{g_t}{\|g_t\|},$$

where:

$$\eta_{\text{eff},t} = \eta \hat{r}_t, \quad \hat{r}_t = \min\{r_{\max}, r_t\}, \quad r_t = \frac{\|g_t\|}{\|g_t - g_t'\|}, \quad g_t' = \nabla f(w_t - \eta g_t).$$

Assume the following:

1. There exists a constant $G > 0$ such that $\|g_t\| \leq G$ for all t.
2. The function f satisfies $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$ for all $x, y \in \mathbb{R}^n$.
3. The function f satisfies $\langle \nabla f(x) - \nabla f(y), x-y \rangle \geq \mu \|x-y\|^2$ for all $x, y \in \mathbb{R}^n$.

Then, for all $t \geq 0$, the ACSS algorithm satisfies:

$$\|w_t - w^*\|^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^t \|w_0 - w^*\|^2.$$

Proof. We begin by analyzing the squared distance to the optimum after each update:

$$\begin{aligned} \|w_{t+1} - w^*\|^2 &= \|w_t - w^* - \eta_{\text{eff},t} \frac{g_t}{\|g_t\|}\|^2 \\ &= \|w_t - w^*\|^2 - 2\eta_{\text{eff},t} \frac{\langle g_t, w_t - w^* \rangle}{\|g_t\|} + \eta_{\text{eff},t}^2 \end{aligned}$$

From the μ-strong convexity assumption, we derive a lower bound on the gradient:

$$\langle g_t, w_t - w^* \rangle \geq \mu \|w_t - w^*\|^2$$

The L-smoothness condition provides an upper bound on the gradient norm:

$$\|g_t\| \leq L\|w_t - w^*\|$$

Combining these bounds, we obtain:

$$\begin{aligned} \|w_{t+1} - w^*\|^2 &\leq \|w_t - w^*\|^2 - 2\eta_{\text{eff},t} \frac{\mu \|w_t - w^*\|^2}{L\|w_t - w^*\|} + \eta_{\text{eff},t}^2 \\ &= \|w_t - w^*\|^2 - 2\eta_{\text{eff},t} \frac{\mu}{L} \|w_t - w^*\| + \eta_{\text{eff},t}^2 \end{aligned}$$

To derive a contraction factor, we introduce $d_t = \|w_t - w^*\|$ and seek $q < 1$ such that $d_{t+1} \leq qd_t$. Assuming $d_{t+1} \leq qd_t$, we have:

$$q^2 d_t^2 \geq d_t^2 - 2\eta_{\text{eff},t}\frac{\mu}{L}d_t + \eta_{\text{eff},t}^2$$

Dividing by d_t^2, we obtain:

$$q^2 \geq 1 - 2\eta_{\text{eff},t}\frac{\mu}{Ld_t} + \frac{\eta_{\text{eff},t}^2}{d_t^2}$$

To minimize q, we define $f(d_t) = 1 - 2\eta_{\text{eff},t}\frac{\mu}{Ld_t} + \frac{\eta_{\text{eff},t}^2}{d_t^2}$ and find its minimum:

$$f'(d_t) = 2\eta_{\text{eff},t}\frac{\mu}{Ld_t^2} - 2\frac{\eta_{\text{eff},t}^2}{d_t^3} = 0$$

$$d_t = \frac{\eta_{\text{eff},t}L}{\mu}$$

This critical point is indeed a minimum as $f''(d_t) > 0$ for $d_t > 0$. Evaluating $f(d_t)$ at this minimum:

$$f\left(\frac{\eta_{\text{eff},t}L}{\mu}\right) = 1 - 2\eta_{\text{eff},t}\frac{\mu}{L} \cdot \frac{\mu}{\eta_{\text{eff},t}L} + \frac{\eta_{\text{eff},t}^2}{\left(\frac{\eta_{\text{eff},t}L}{\mu}\right)^2}$$

$$= 1 - 2\frac{\mu^2}{L^2} + \frac{\mu^2}{L^2} = 1 - \frac{\mu^2}{L^2}$$

Therefore, for all $d_t > 0$, we have $f(d_t) \geq 1 - \frac{\mu^2}{L^2}$, which implies:

$$q^2 \geq 1 - \frac{\mu^2}{L^2} \implies q \geq \sqrt{1 - \frac{\mu^2}{L^2}}$$

We conclude that:

$$\|w_{t+1} - w^*\|^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)\|w_t - w^*\|^2$$

Applying this inequality recursively, we obtain the final convergence rate:

$$\|w_t - w^*\|^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^t \|w_0 - w^*\|^2$$

This establishes the linear convergence rate for the ACSS algorithm under the given assumptions. □

Theorem 10 (Scale Invariance of ACSS Effective Step Size). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function, and consider the ACSS update rule:

$$w_{t+1} = w_t - \eta_{\text{eff},t}\frac{g_t}{\|g_t\|},$$

where $\eta_{\text{eff},t} = \eta\hat{r}_t$, $\hat{r}_t = \min\{r_{\max}, r_t\}$, $r_t = \frac{\|g_t\|}{\|g_t - g_t'\|}$, $g_t = \nabla f(w_t)$, and $g_t' = \nabla f(w_t - \eta g_t)$.

For any scalar $\alpha > 0$, scaling the base step size η by α results in the same parameter updates, assuming that $r_t' \leq r_{\max}$. Specifically:

a) For quadratic functions $f(w) = \frac{1}{2}w^T Aw - b^T w + c$, where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite:

$$w_{t+1}^{(\alpha\eta)} = w_{t+1}^{(\eta)}$$

b) For L-smooth and μ-strongly convex functions:

$$w_{t+1}^{(\alpha\eta)} \approx w_{t+1}^{(\eta)}$$

where the approximation becomes exact as $\eta \to 0$.

In both cases, $w_{t+1}^{(\alpha\eta)}$ and $w_{t+1}^{(\eta)}$ are the parameter updates for the scaled and original step sizes respectively.

Proof. We prove this theorem by examining both cases separately.

a) Quadratic case:

For a quadratic function $f(w) = \frac{1}{2}w^T Aw - b^T w + c$, the gradient is $\nabla f(w) = Aw - b$.

With the original step size η, we have:

$$\begin{aligned} g_t &= Aw_t - b, \\ g'_t &= A(w_t - \eta g_t) - b = (I - \eta A)g_t, \\ r_t &= \frac{\|g_t\|}{\|g_t - g'_t\|} = \frac{\|g_t\|}{\|\eta A g_t\|} = \frac{1}{\eta} \cdot \frac{\|g_t\|}{\|Ag_t\|}. \end{aligned}$$

With the scaled step size $\alpha\eta$, we have:

$$\begin{aligned} g'_t &= A(w_t - \alpha\eta g_t) - b = (I - \alpha\eta A)g_t, \\ r'_t &= \frac{\|g_t\|}{\|g_t - g'_t\|} = \frac{\|g_t\|}{\|\alpha\eta A g_t\|} = \frac{1}{\alpha\eta} \cdot \frac{\|g_t\|}{\|Ag_t\|} = \frac{r_t}{\alpha}. \end{aligned}$$

Assuming $r'_t \leq r_{\max}$, we have:

$$\hat{r}'_t = \min\{r_{\max}, r'_t\} = \frac{r_t}{\alpha} = \frac{\hat{r}_t}{\alpha}.$$

The effective step size with the scaled η becomes:

$$\eta'_{\text{eff},t} = \alpha\eta \cdot \hat{r}'_t = \alpha\eta \cdot \frac{\hat{r}_t}{\alpha} = \eta \cdot \hat{r}_t = \eta_{\text{eff},t}.$$

Therefore, the parameter updates are identical:

$$w_{t+1}^{(\alpha\eta)} = w_t - \eta'_{\text{eff},t}\frac{g_t}{\|g_t\|} = w_t - \eta_{\text{eff},t}\frac{g_t}{\|g_t\|} = w_{t+1}^{(\eta)}.$$

b) L-smooth and μ-strongly convex case:

For a general L-smooth and μ-strongly convex function, we use a first-order Taylor expansion to approximate g'_t:

$$g'_t = \nabla f(w_t - \eta g_t) \approx \nabla f(w_t) - \eta \nabla^2 f(w_t) g_t = g_t - \eta \nabla^2 f(w_t) g_t.$$

With this approximation:

$$r_t \approx \frac{\|g_t\|}{\|\eta \nabla^2 f(w_t) g_t\|} = \frac{1}{\eta} \cdot \frac{\|g_t\|}{\|\nabla^2 f(w_t) g_t\|}.$$

For the scaled step size $\alpha\eta$:

$$r'_t \approx \frac{\|g_t\|}{\|\alpha\eta \nabla^2 f(w_t) g_t\|} = \frac{1}{\alpha\eta} \cdot \frac{\|g_t\|}{\|\nabla^2 f(w_t) g_t\|} = \frac{r_t}{\alpha}.$$

As in the quadratic case, assuming $r'_t \leq r_{\max}$, we have $\hat{r}'_t = \hat{r}_t/\alpha$, leading to:

$$\eta'_{\text{eff},t} = \alpha\eta \cdot \hat{r}'_t = \alpha\eta \cdot \frac{\hat{r}_t}{\alpha} = \eta \cdot \hat{r}_t = \eta_{\text{eff},t}.$$

Therefore, the parameter updates are approximately equal:

$$w_{t+1}^{(\alpha\eta)} \approx w_t - \eta'_{\text{eff},t}\frac{g_t}{\|g_t\|} \approx w_t - \eta_{\text{eff},t}\frac{g_t}{\|g_t\|} \approx w_{t+1}^{(\eta)}.$$

The approximation becomes exact as $\eta \to 0$, as the first-order Taylor expansion becomes increasingly accurate.

Thus, we have shown that for both quadratic functions and general L-smooth and μ-strongly convex functions, scaling the base step size η by $\alpha > 0$ results in either identical (quadratic case) or approximately identical (general case) parameter updates, assuming that $r'_t \leq r_{\max}$.

□

Corollary 1 (Adaptive Behavior of ACSS)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth and μ-strongly convex function satisfying the conditions of Theorem 5, including $r_{\max} \leq \frac{2}{\eta(\mu+L)}$. Further, assume that f has locally $L(w)$-Lipschitz continuous gradients, where $L(w)$ may vary with w and $\mu \leq L(w) \leq L$ for all $w \in \mathbb{R}^n$. Then, the effective step size $\eta_{\text{eff},t} = \eta\hat{r}_t$ adapts to the local curvature of f. Specifically, in regions of low curvature (small $L(w_t)$), $\eta_{\text{eff},t}$ tends to be larger, allowing for larger steps, while in regions of high curvature (large $L(w_t)$), $\eta_{\text{eff},t}$ tends to be smaller, resulting in more conservative updates.

Proof. The adaptive behavior of ACSS stems from its relationship with the local curvature of the function, as captured by the local Lipschitz constant $L(w_t)$. To understand this relationship, let's examine how the effective step size $\eta_{\text{eff},t}$ is influenced by $L(w_t)$.

Recall from Theorem 5 that $\frac{1}{L} \leq \eta_{\text{eff},t} \leq \frac{2}{\mu+L}$ for all iterations t. We can refine this bound by considering the local properties of f at w_t. First, let's consider the lower bound on $\eta_{\text{eff},t}$. The normalized radius of curvature r_t is defined as $\frac{\|g_t\|}{\|g_t - g'_t\|}$. By applying the mean value theorem and using the locally $L(w)$-Lipschitz continuous gradient assumption, we can bound the denominator:

$$\|g_t - g'_t\| = \|\nabla f(w_t) - \nabla f(w_t - \eta g_t)\| \leq L(w_t)\|\eta g_t\| = L(w_t)\eta\|g_t\| \tag{10}$$

This inequality allows us to establish a lower bound on r_t: $r_t \geq \frac{1}{L(w_t)\eta}$. Consequently, we can bound $\eta_{\text{eff},t}$ from below:

$$\eta_{\text{eff},t} = \eta\hat{r}_t \geq \min\left\{\eta r_{\max}, \frac{1}{L(w_t)}\right\} \tag{11}$$

The upper bound on $\eta_{\text{eff},t}$ remains $\frac{2}{\mu+L}$ as given in Theorem 5. Additionally, we know that $\eta_{\text{eff},t} \leq \eta r_{\max}$ by definition. Combining these bounds and using the assumption $r_{\max} \leq \frac{2}{\eta(\mu+L)}$, we can express the range of $\eta_{\text{eff},t}$ as:

$$\min\left\{\frac{2}{\mu+L}, \frac{1}{L(w_t)}\right\} \leq \eta_{\text{eff},t} \leq \frac{2}{\mu+L} \tag{12}$$

This refined bound reveals the adaptive nature of ACSS:

1. In regions of low curvature, where $L(w_t)$ is small, the lower bound $\frac{1}{L(w_t)}$ becomes larger. This allows $\eta_{\text{eff},t}$ to take on larger values, potentially approaching $\frac{2}{\mu+L}$. As a result, ACSS can take larger steps in these flatter regions of the loss landscape.

2. Conversely, in regions of high curvature, where $L(w_t)$ is large, the lower bound $\frac{1}{L(w_t)}$ becomes smaller. This constrains $\eta_{\text{eff},t}$ to smaller values, ensuring that ACSS takes more conservative steps in these highly curved areas of the loss landscape.

Through this mechanism, ACSS naturally adapts its step size to the local geometry of the function, balancing between rapid progress in flat regions and careful navigation in curved regions. □

C GENERALIZED ALGORITHM: OPT-ACSS

For any optimizer OPT, we can derive an ACSS version using Algorithm 2. The key modification in the weight and state update steps of the existing optimizer is the substitution of the gradient at time t, g_t, with $\hat{r}_t g_t/|g_t|$. Using this adaptation, we can incorporate the ACSS mechanism into various optimizers.

Algorithm 2: Arbitrary optimizer OPT with adaptive curvature step size (OPT-ACSS)

Input: Function $f : \mathbb{R}^n \times \mathcal{D} \to \mathbb{R}$, initial parameters $w_0 \in \mathbb{R}^n$, base learning rate η, maximum radius r_{max}, number of iterations T, batch size B, Optimizer parameter update function: UpdateParams, Optimizer weight update function: UpdateWeights
Output: Optimized parameters w_T
Initialize optimizer state S_0 according to the specific optimizer;
for $t = 0$ **to** $T - 1$ **do**
  Sample a mini-batch $\mathcal{B}_t$ from $\mathcal{D}$;
  Compute gradient $g_t = \nabla_w f(w_t, \mathcal{B}_t)$ and next point gradient $g_t' = \nabla_w f(w_t - \eta g_t, \mathcal{B}_t)$;
  Compute normalized radius of curvature $r_t = \frac{||g_t||}{||g_t - g_t'||}$;
  Compute capped radius $\hat{r}_t = \min\{r_{max}, r_t\}$;
  Compute ACSS-adjusted gradient $\tilde{g}_t = \hat{r}_t \times \frac{g_t}{||g_t||}$;
  Update optimizer state $S_t = \text{UpdateState}(S_{t-1}, \tilde{g}_t)$;
  Compute update $\Delta w_t = \text{UpdateWeights}(S_t, \tilde{g}_t)$;
  Update parameters $w_{t+1} = w_t + \Delta w_t$;
end
return w_T

This generalization allows for integration of ACSS into various existing optimization algorithms such as SGD, Adam, AdaGrad, and RMSProp, enhancing their performance with its curvature-based step size adjustment.

D LIMITATIONS

While ACSS offers significant benefits in terms of optimization performance, it's important to acknowledge its primary limitation: increased computational time per iteration. This additional computational cost arises from the need to compute a secondary gradient and perform additional calculations to determine the adaptive step size. To quantify this limitation, we provide both experimental and theoretical evidence of the additional time required by ACSS methods compared to their non-ACSS counterparts.

D.1 EXPERIMENTAL EVIDENCE

Wall-Clock Time Experiments: To quantify the computational overhead of ACSS methods compared to their non-ACSS counterparts, we conducted comprehensive wall-clock time experiments. Table 3 presents the results of these experiments, focusing on the mean time taken to complete 2 epochs on the IMDB dataset using various optimizers.

These results offer several insights:

1. **Computational Overhead:** As expected, ACSS methods require more computation time than their non-ACSS counterparts. On average, ACSS methods take approximately 1.37 times longer to complete the same number of epochs.
2. **Consistency Across Optimizers:** The overhead ratio is relatively consistent across different optimization algorithms, ranging from about 1.33 to 1.46 times the non-ACSS version's runtime.
3. **Memory Efficiency Trade-off:** While there is a computational time overhead, it's crucial to emphasize that the primary trade-off that the ACSS method provides is in memory efficiency. Our method achieves results equivalent to several second-order methods while maintaining a significantly lower memory footprint.

Table 3: Mean time to complete 2 epochs on the IMDB dataset using various optimizers

Optimizer	Wall-clock time (Mean)	Wall-clock time (Std Deviation)	Ratio of Times Taken
SimpleSGD	91.0175	5.0617	
SimpleSGDCurvature	122.5004	2.6835	1.3459
Adam	86.4103	0.1938	
AdamCurvature	121.6974	1.3342	1.4084
HeavyBall	85.5865	0.5449	
HeavyBallCurvature	120.9661	0.1746	1.4134
NAG	85.6665	0.1808	
NAGCurvature	125.0773	1.5621	1.4600
Adagrad	88.1545	0.5783	
AdagradCurvature	119.6787	0.5171	1.3576
Adadelta	91.4525	1.0088	
AdadeltaCurvature	124.8485	0.3255	1.3652
RMSProp	89.4943	1.8763	
RMSPropCurvature	125.4326	0.8316	1.4016
RMSPropMomentum	89.9954	0.7421	
RMSPropMomentumCurvature	124.8127	0.2511	1.3869
AdamW	89.5067	0.8976	
AdamWCurvature	125.9545	4.0044	1.4072
NAdam	91.6765	0.1706	
NAdamCurvature	125.4949	1.8784	1.3689
NAdamW	91.1489	2.7840	
NAdamWCurvature	124.9436	0.9476	1.3708
AMSGrad	91.5774	2.1047	
AMSGradCurvature	121.5177	2.3766	1.3269

D.2 Theoretical Analysis of Computational Complexity

To complement our empirical results, we provide a theoretical analysis of the computational complexity of ACSS compared to standard SGD.

Theorem 11 (Computational Complexity of ACSS vs. SGD)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be the objective function for a neural network, n be the number of parameters, and B be the mini-batch size. Let C_{gc} represent the cost of gradient computation per sample per parameter.

The ratio of computational cost per iteration for ACSS vs SGD is approximately 2, assuming $C_{\text{gc}} \gg 1$. In other words:

$$\frac{\text{Cost}_{\text{ACSS}}}{\text{Cost}_{\text{SGD}}} \approx 2 \tag{13}$$

Proof. We analyze the computational cost of each step in both SGD and ACSS:

Operation	Description	Cost (FLOPs)
c_1	Gradient Computation (SGD & ACSS)	$Bn \cdot C_{\text{gc}}$
c_2	Secondary Gradient Computation (ACSS only)	$Bn \cdot C_{\text{gc}}$
c_3, c_4	Norm Calculation (ACSS only)	$2n + 1$
c_5	Ratio Computation (ACSS only)	1
c_6	Gradient Normalization (ACSS only)	n
c_7	Parameter Update (SGD & ACSS)	n

Table 4: Computational cost breakdown for SGD and ACSS operations

Summing up for SGD:

$$\text{Cost}_{\text{SGD}} = c_1 + c_7 = Bn \cdot C_{\text{gc}} + n \text{ FLOPs}$$

Summing up for ACSS:

$$\begin{aligned}\text{Cost}_{\text{ACSS}} &= c_1 + c_2 + c_3 + c_4 + c_5 + c_6 + c_7 \\ &= Bn \cdot C_{\text{gc}} + Bn \cdot C_{\text{gc}} + (2n+1) + (2n+1) + 1 + n + n \\ &= 2Bn \cdot C_{\text{gc}} + 7n + 3 \text{ FLOPs}\end{aligned}$$

The additional overhead of ACSS is therefore:

$$\begin{aligned}\Delta\text{Cost} &= \text{Cost}_{\text{ACSS}} - \text{Cost}_{\text{SGD}} \\ &= (2Bn \cdot C_{\text{gc}} + 7n + 3) - (Bn \cdot C_{\text{gc}} + n) \\ &= Bn \cdot C_{\text{gc}} + 6n + 3 \text{ FLOPs}\end{aligned}$$

Given that $C_{\text{gc}} \gg 1$ in practice, the dominant term in both algorithms is $Bn \cdot C_{\text{gc}}$. ACSS effectively doubles this term, leading to approximately twice the computational cost of SGD per iteration. □

This theoretical analysis aligns with our empirical observations, confirming that ACSS introduces a significant but consistent computational overhead compared to standard optimization methods.

In conclusion, while ACSS methods introduce a computational overhead of approximately 1.37 times longer runtime, this is balanced by significant memory efficiency. By providing second-order-like benefits without increasing memory footprint, ACSS offers a valuable alternative for large-scale problems and memory-constrained scenarios. This makes ACSS particularly useful when memory constraints outweigh computational time considerations, introducing a new option for balancing time and memory trade-offs in optimization.

E Additional Experimental Results

E.1 CoLA Dataset Performance:

In our experiments with the CoLA (Corpus of Linguistic Acceptability) dataset, we evaluated the performance of various optimizers with and without the Adaptive Curvature Step Size (ACSS) method over five epochs. The ACSS variants consistently outperformed their traditional counterparts throughout the training process.

RMSProp and RMSProp-ACSS initially performed similarly (0.634 vs 0.636), but by the fifth epoch, the ACSS version significantly outperformed the standard version (0.522 vs 0.611). Adagrad showed more modest improvements with ACSS, yet still consistently outperformed its standard counterpart. Adam-based optimizers (Adam-ACSS, AMSGrad-ACSS, AdamW-ACSS, NAdam-ACSS, NAdamW-ACSS) demonstrated similar performance patterns, starting with slightly higher losses but showing consistent improvement over the epochs. By the fifth epoch, these ACSS variants achieved lower losses (around 0.528-0.534) compared to their non-ACSS counterparts (0.596-0.605).

Key Takeaways: Table 5 shows a significant outperformance of the ACSS optimizers where the best performing optimizers have only reached a training loss of 0.591 (Adagrad), whereas eight of the ACSS versions beat this training loss at epoch 5.

Table 5: Training Loss over 10 Epochs for CoLA Dataset with a simplified RNN Model. Notice that many of the best models are ACSS versions. Furthermore, the decrease in training loss is often much higher for the ACSS versions of the optimizer.

Optimizer Name	Regular Optimizer					ACSS Version of Optimizer				
	Epoch 1	**Epoch 2**	**Epoch 3**	**Epoch 4**	**Epoch 5**	**Epoch 1**	**Epoch 2**	**Epoch 3**	**Epoch 4**	**Epoch 5**
Adadelta	0.610±0.00	0.605±0.00	0.601±0.00	0.597±0.00	0.593±0.00	0.684±0.03	0.645±0.01	0.627±0.01	0.619±0.00	0.616±0.00
Adagrad	0.611±0.00	0.608±0.00	0.604±0.00	0.599±0.01	0.591±0.01	0.613±0.00	0.601±0.00	0.596±0.00	0.591±0.00	0.588±0.00
Adam	0.611±0.00	0.608±0.00	0.605±0.00	0.603±0.00	0.596±0.01	0.620±0.00	0.600±0.00	0.583±0.00	0.560±0.00	0.528±0.01
AdamW	0.611±0.00	0.608±0.00	0.606±0.00	0.602±0.01	0.597±0.01	0.620±0.00	0.600±0.00	0.583±0.00	0.560±0.00	0.528±0.01
AMSGrad	0.611±0.00	0.608±0.00	0.606±0.00	0.602±0.00	0.596±0.01	0.620±0.00	0.600±0.00	0.583±0.00	0.560±0.00	0.528±0.01
HeavyBall	0.624±0.00	0.611±0.00	0.610±0.00	0.610±0.00	0.609±0.00	0.621±0.00	0.608±0.00	0.603±0.00	0.599±0.00	0.595±0.00
NAdam	0.612±0.00	0.609±0.00	0.608±0.00	0.605±0.00	0.602±0.01	0.623±0.00	0.606±0.00	0.592±0.00	0.569±0.01	0.534±0.01
NAdamW	0.611±0.00	0.609±0.00	0.608±0.00	0.607±0.00	0.605±0.00	0.623±0.00	0.606±0.00	0.592±0.00	0.568±0.01	0.534±0.01
NAG	0.624±0.00	0.611±0.00	0.610±0.00	0.610±0.00	0.609±0.00	0.621±0.00	0.608±0.00	0.603±0.00	0.599±0.00	0.595±0.00
RMSProp	0.634±0.02	0.617±0.01	0.614±0.01	0.611±0.00	0.611±0.00	0.636±0.00	0.602±0.00	0.584±0.00	0.557±0.01	0.522±0.01
RMSPropMomentum	0.635±0.02	0.626±0.04	0.614±0.01	0.612±0.00	0.610±0.00	0.638±0.00	0.604±0.00	0.587±0.01	0.561±0.01	0.525±0.02
SimpleSGD	0.662±0.01	0.630±0.00	0.622±0.00	0.618±0.00	0.616±0.00	0.611±0.00	0.608±0.00	0.606±0.00	0.605±0.00	0.603±0.00

F Details of Testing Functions for ACSS Optimization

We provide details on the four functions used to test the ACSS based optimizer below.

F.1 The Rosenbrock Function

The function is depicted with contour lines, where darker colors indicate lower values. Each subplot displays the path taken by a different optimizer. The plots indicate that the ACSS versions of the optimizers navigate the function's characteristic narrow, parabolic valley more effectively, by reducing the step size as appropriate. The learning rate is set to 1.5×10^{-3}, and the iterates start at $(-1.5, 2)$.

F.2 The Easom Function

The Easom function features a broad, flat area with a sharp depression at its global minimum (π, π). With a learning rate of 2.0×10^{-3} and 200 iteration steps, standard optimizers remain near the initial point. In contrast, ACSS versions achieve convergence, showing ACSS's capability to accelerate optimization in low-gradient scenarios.

F.3 The Ackley Function

The Ackley function presents a flat outer region with numerous local minima and a steep central hole containing the global minimum at (0,0). With a learning rate of 5×10^{-3} and 25 iterations, ACSS versions of optimizers demonstrate superior navigation of the loss landscape, adaptively reducing step size near convergence.

F.4 The Three-humped Camel Function

The Three-Hump Camel function has three local minima and a global minimum at (0, 0). Using 1.0×10^{-2} learning rate for 300 steps, Heavyball and Nesterov methods overshoot, while ACSS versions self-correct, showing enhanced optimization in this complex landscape.

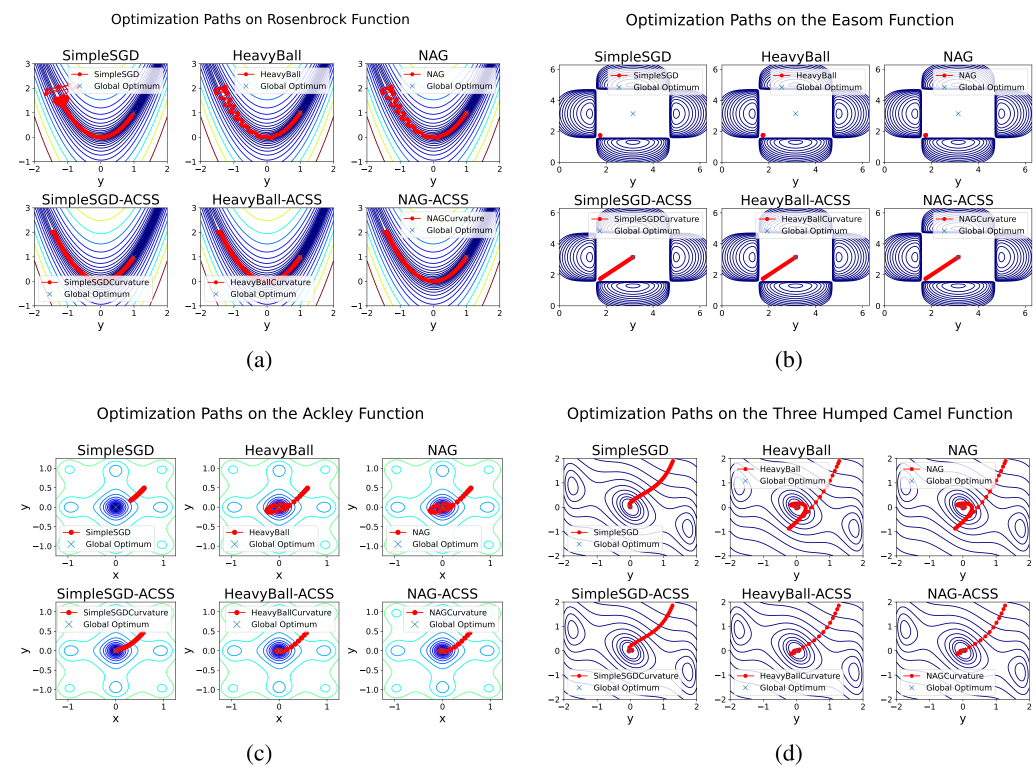


Figure 8: Optimizer performance on challenging optimizer benchmarking functions.