CURVATURE-SCALED SHARPNESS-AWARE MINIMIZATION: CONVERGENCE BEYOND SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sharpness-Aware Minimization (SAM) has established itself as a widely used method for training deep neural networks. Despite its popularity, existing convergence guarantees for SAM-type methods require smoothness, a restrictive assumption. Generalized smoothness conditions have recently been proposed in the analysis of gradient methods but remain underexplored for SAM. To analyze the effect of SAM's perturbation under generalized smoothness, we use relative gradient distortion to quantify how the sharpness-aware direction departs from the original gradient. Under (L_0, L_1)-smoothness, we show that this distortion can become unbounded near stationarity for SAM and at large gradients for its unnormalized variant (USAM), with examples where SAM cycles and USAM diverges. We then establish a necessary condition for uniform distortion control among perturbations along the gradient whose nonnegative multipliers depend only on its norm. Motivated by this condition, we propose Curvature-Scaled SAM (CS-SAM), which scales both the perturbation and the descent step by suitable curvature upper bounds, and prove that it uniformly bounds the distortion. Under both (L_0, L_1)- and (H_0, H_1)-smoothness, we establish deterministic $\mathcal{O}(1/T)$ rates for the best squared gradient norm on nonconvex objectives, last-iterate $\mathcal{O}(1/T)$ rates for the function gap on convex objectives with a minimizer, and linear convergence under the Polyak–Łojasiewicz condition. With zero perturbation, our guarantees recover the rates for gradient descent with (L_0, L_1)- or (H_0, H_1)-based step-sizes. Minibatch experiments on computer vision tasks demonstrate competitive performance for both variants, with the (H_0, H_1) variant achieving the highest mean best-test accuracy in all four settings against tuned SAM and USAM with constant and cosine annealing schedules.

1 INTRODUCTION

Sharpness-Aware Minimization (SAM) was introduced to improve generalization by seeking parameters whose training loss remains low throughout a local neighborhood (Foret et al., 2021). It pursues this objective through a simple two-step procedure: first perturb the parameters in a direction of increasing loss, then update the original parameters using the gradient evaluated at the perturbed point. The choice of perturbation therefore determines how the resulting update departs from ordinary gradient descent. Standard SAM uses a fixed-radius perturbation of length r along the normalized gradient, whereas its unnormalized variant (USAM) uses a fixed multiple ρ of the gradient itself (Andriushchenko & Flammarion, 2022):

$$y_t = x_t + r\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}, \qquad x_{t+1} = x_t - \eta_t \nabla f(y_t), \tag{SAM}$$

$$y_t = x_t + \rho \nabla f(x_t), \qquad x_{t+1} = x_t - \eta_t \nabla f(y_t), \tag{USAM}$$

with radius $r > 0$, multiplier $\rho \geq 0$, and outer step $\eta_t > 0$. Both rules cost two gradient evaluations per iteration, one at the iterate x_t and one at the perturbed point y_t.

Throughout, we consider the problem

$$\min_{x \in \mathbb{R}^d} f(x), \qquad f_\star := \inf_{x \in \mathbb{R}^d} f(x) > -\infty, \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable. In model training, x represents the model parameters and f the training loss. We call the Hessian operator norm $\left\|\nabla^2 f(x)\right\|$ the local curvature at x. Classical smoothness bounds this quantity by a constant $L > 0$:

$$\left\|\nabla^2 f(x)\right\| \leq L \qquad \forall x \in \mathbb{R}^d. \tag{2}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Classical L-smoothness can fail even for simple neural-network training objectives (Alimisis et al., 2026). Motivated by empirical observations that local curvature grows with the gradient norm during neural-network training, Zhang et al. (2020b) introduced (L_0, L_1)-smoothness, which allows the Hessian norm to grow with the gradient norm. Subsequent theoretical results support this condition for specific neural-network models. In particular, the bounds of Taheri & Thrampoulidis (2023) imply (L_0, L_1)-smoothness for two-layer networks with fixed output weights and smoothed leaky-ReLU activations, trained with exponential loss on linearly separable data. Complementary results establish (H_0, H_1)-smoothness for two-layer networks trained with ℓ_2-regularized cross-entropy loss under suitable activation assumptions (Alimisis et al., 2026). These findings motivate the gradient-dependent and loss-dependent curvature bounds considered here:

$$\left\|\nabla^2 f(x)\right\| \le M_L(x) := L_0 + L_1 \left\|\nabla f(x)\right\|, \qquad (L_0, L_1)\text{-smoothness}, \tag{3}$$

$$\left\|\nabla^2 f(x)\right\| \le M_H(x) := H_0 + H_1\big(f(x) - f_\star\big), \qquad (H_0, H_1)\text{-smoothness}, \tag{4}$$

for every $x \in \mathbb{R}^d$, with $L_0 > 0$, $L_1 \ge 0$, and $H_0, H_1 \ge 0$. Note that both models contain classical smoothness as a special case ($L_1 = 0$ or $H_1 = 0$). With slight abuse of notation, in several parts of the paper we write $M_L := M_L(x)$ and $M_H := M_H(x)$. We refer to a positive function that upper-bounds the local curvature everywhere as a *curvature scale*; M_L and, for nonconstant objectives, M_H provide the two scales studied in this work (Proposition 2.4).

The difficulty is that the curvature bound must control the gradient change along the perturbation, as well as the subsequent descent step. Even under classical smoothness, a fixed perturbation radius can prevent exact convergence: Bartlett et al. (2023) analyzed cycling of fixed-radius SAM on convex quadratics, and Si & Yun (2023) established stationarity neighborhoods in several settings. The inexact-gradient viewpoint of Khanh et al. (2024) makes the perturbation-induced error explicit, while the clipped rule of Adly & Le (2026) controls it near stationarity by making the perturbation proportional to the gradient. Under (3), however, a gradient-proportional perturbation can also become too large when the gradient grows. Thus, extending SAM to generalized smoothness requires control at both ends of the gradient scale.

For ordinary gradient methods, adapting to curvature growth is already well studied. The (L_0, L_1) model was introduced by Zhang et al. (2020b) to explain gradient clipping and has since supported analyses of clipped, normalized, and adaptive methods (Zhang et al., 2020a; Crawshaw et al., 2022; Chen et al., 2023; Koloskova et al., 2023; Li et al., 2023). In particular, Gorbunov et al. (2025) and Vankov et al. (2025) analyzed steps inversely proportional to gradient-dependent curvature bounds. The loss-dependent model has likewise been used to explain learning-rate warm-up (Alimisis et al., 2026) and to analyze line search and adaptive gradient methods (Vaswani & Babanezhad Harikandeh, 2025; Vaswani et al., 2026). These results motivate scaling the outer descent step by a curvature bound. The remaining question for SAM is how that bound should also determine the perturbation.

1.1 Main Contributions

◇ **Complementary failures and a necessary condition for distortion control.** In Section 2.1, we show that the relative gradient distortion of SAM can become unbounded as the gradient vanishes, while that of USAM can become unbounded as the gradient grows. Proposition 2.1 gives explicit examples within the (L_0, L_1)-smooth class: fixed-radius SAM admits a nonzero two-cycle on a quadratic, and USAM diverges from sufficiently large initializations on a quartic. Motivated by these complementary failures, Theorem 2.2 establishes a necessary condition for uniformly bounded distortion over this class among perturbations along the gradient whose nonnegative multipliers depend only on its norm. The perturbation length must be at most proportional to the gradient norm near stationarity and bounded at large gradients. This condition motivates the gradient-dependent perturbation of CS-SAM.

◇ **Curvature-Scaled Sharpness-Aware Minimization.** In Section 2.2, we construct CS-SAM using the perturbation multiplier $\rho/M(x_t)$ and the descent step $\gamma/M(y_t)$, where $\rho \ge 0$ and $\gamma > 0$. We identify conditions on the curvature scale M that ensure uniform distortion control and show that both M_L and M_H satisfy them. For every scale satisfying these conditions, Theorem 2.5 bounds the distortion by $e^\rho - 1$ at every nonstationary iterate.

◇ **Convergence under generalized smoothness.** We establish convergence guarantees for CS-SAM under (L_0, L_1)-smoothness with $M = M_L$ and under (H_0, H_1)-smoothness with $M = M_H$ (Theorems 3.1 and 3.2). In each setting, for suitable perturbation and step parameters, we prove

Table 1: Comparison with the closest prior work. $D(x,d) = \|\nabla f(x+d) - \nabla f(x)\| / \|\nabla f(x)\|$ is the relative gradient distortion of the perturbation d (Section 2.1). Curvature models: L means $\|\nabla^2 f(x)\| \le L$; (L_0, L_1) means $\|\nabla^2 f(x)\| \le L_0 + L_1 \|\nabla f(x)\|$; (H_0, H_1) means $\|\nabla^2 f(x)\| \le H_0 + H_1(f(x) - f_\star)$, with $f_\star = \inf_x f(x)$. ✓: D is bounded by a constant; ✗: the bound degenerates. Rates are on the best squared gradient norm over T iterations (nonconvex), on the last-iterate gap $f(x_T) - f_\star$ for convex objectives with a minimizer (convex), and on the function gap under PL.

Method	Perturbation $d(x)$	Curvature model	D uniformly bounded	Nonconvex	Convex	PL
SAM (Foret et al., 2021) rates: (Si & Yun, 2023), (Oikonomou & Loizou, 2025)	$r\,\nabla f(x)/\|\nabla f(x)\|$	L	✗	$\mathcal{O}(1/T) + L^2 r^2$	no last-iterate gap rate	linear to a neighborhood
USAM (Andriushchenko & Flammarion, 2022)	$\rho \nabla f(x)$	L	✓	$\mathcal{O}(1/T)$	$\mathcal{O}(1/T)$	linear
Unified SAM (Oikonomou & Loizou, 2025)	$\rho[(1-\lambda) + \lambda/\|\nabla f(x)\|] \cdot \nabla f(x)$	L	✓ if $\lambda = 0$; ✗ otherwise	$\mathcal{O}(T^{-1/2})$ (in expectation)	not analyzed	$\lambda = 0$: linear $\lambda > 0$: linear to a neighborhood
Clipped perturbation (Adly & Le, 2026)	$\rho_k \nabla f(x)/\|\nabla f(x)\|^\alpha$, $\rho_k = \min\{\rho, \frac{\theta}{L}\|\nabla f(x)\|^\alpha\}$	L	✓	$\mathcal{O}(1/T)$	not analyzed	not analyzed
Inexact-gradient SAM (Khanh et al., 2024)	normalized and unnormalized variants	L	normalized: ✗ USAM: ✓	$\|\nabla f(x_t)\| \to 0$ (no generic rate)	$f(x_T) \to f_\star$ (no rate)	USAM: linear under KL exponent $1/2$
Ours, CS-SAM with M_L	$\rho \nabla f(x)/(L_0 + L_1\|\nabla f(x)\|)$	(L_0, L_1)	✓	$\mathcal{O}(1/T)$	$\mathcal{O}(1/T)$	linear
Ours, CS-SAM with M_H	$\rho \nabla f(x)/(H_0 + H_1(f(x) - f_\star))$	(H_0, H_1)	✓	$\mathcal{O}(1/T)$	$\mathcal{O}(1/T)$	linear

$\mathcal{O}(1/T)$ rates for the best squared gradient norm on nonconvex objectives and last-iterate $\mathcal{O}(1/T)$ rates for the function gap on convex objectives with a minimizer. We also prove linear convergence in both settings under the Polyak–Łojasiewicz (PL) condition, $\|\nabla f(x)\|^2 / 2 \ge \mu(f(x) - f_\star)$ for every x and some $\mu > 0$.

◇ **Recovery of gradient descent guarantees for $\rho = 0$.** Setting $\rho = 0$ recovers gradient descent with step $\gamma/M(x_t)$. Our bounds recover the known nonconvex and convex rates for gradient-dependent steps (Vankov et al., 2025; Gorbunov et al., 2025), and the convex and PL rates for loss-dependent steps (Alimisis et al., 2026), while allowing larger step sizes in the latter case. Thus the analysis extends existing gradient descent guarantees to positive sharpness-aware perturbations.

◇ **Numerical Evaluation.** In Section 4.1, we verify Proposition 2.1 and Theorem 2.5: SAM cycles and USAM diverges, whereas CS-SAM converges with a distortion below $e^\rho - 1$ for both scales. In Section 4.2, we evaluate minibatch variants beyond the deterministic setting of our theory on ResNet-20/32 and CIFAR-10/100. Both use two gradient evaluations per iteration. Against tuned constant-step and cosine annealing-scheduled SAM and USAM, the loss-dependent variant achieves the highest mean best-test accuracy in all four settings, while the gradient-dependent variant exceeds all these baselines in three of four settings.

Table 1 summarizes the distortion control and convergence guarantees alongside prior work.

2 From gradient distortion to curvature-scaled SAM

In this section, we define relative gradient distortion and explain how it controls the difference between a sharpness-aware direction and the gradient used by gradient descent. We then show that SAM and USAM can lose this control at opposite ends of the gradient scale, and derive a necessary condition for keeping it uniformly bounded. This condition motivates the CS-SAM update, for which we prove a uniform distortion bound under both generalized smoothness models. The proofs are deferred to Appendix C; the distortion bound follows from the auxiliary lemmas in Appendix B.

2.1 Relative gradient distortion and the failures of SAM and USAM

The outer step of SAM and USAM is a gradient step that uses $\nabla f(y_t)$ in place of $\nabla f(x_t)$, so a descent analysis must account for the effect of this substitution as well as the step size. We now measure the relative error of this substitution. For $\nabla f(x) \neq 0$ and a displacement d, we define the *relative gradient distortion*

$$D(x,d) := \frac{\|\nabla f(x+d) - \nabla f(x)\|}{\|\nabla f(x)\|}. \tag{5}$$

Let us explain why this is the right quantity. If $D(x,d) \le \delta < 1$, then $\nabla f(x+d)$ makes an acute angle with $\nabla f(x)$, since $\langle \nabla f(x), \nabla f(x+d) \rangle \ge (1-\delta)\|\nabla f(x)\|^2$, and has comparable length, since $(1-\delta)\|\nabla f(x)\| \le \|\nabla f(x+d)\| \le (1+\delta)\|\nabla f(x)\|$. These are exactly the two properties the descent lemma requires, so $D(x,d) \le \delta < 1$ is the property we want a perturbation to have, and

we want it at *every* point. Controlling the perturbation-induced gradient error is also the viewpoint of the inexact-gradient analyses of Khanh et al. (2024), and relative control of this error near stationarity is what the clipped rule of Adly & Le (2026) enforces; the quantity itself is therefore not new. Our question is which perturbation rules keep it uniformly bounded under generalized smoothness.

We first record what the two base rules can guarantee. Under (2) we have

$$D_{\mathrm{SAM}} \le \frac{Lr}{\|\nabla f(x)\|}, \qquad D_{\mathrm{USAM}} \le L\rho. \tag{6}$$

Note that the bound for SAM degenerates as $\|\nabla f(x)\| \to 0$, while the bound for USAM is uniform. Under the more general (3), for $L_1 > 0$ (both bounds are proved in Appendix C):

$$D_{\mathrm{SAM}} \le \left(e^{L_1 r} - 1\right)\left(1 + \frac{L_0}{L_1\|\nabla f(x)\|}\right), \qquad D_{\mathrm{USAM}} \le \frac{L_0 + L_1\|\nabla f(x)\|}{L_1\|\nabla f(x)\|}\left(e^{L_1\rho\|\nabla f(x)\|} - 1\right). \tag{7}$$

The bound for SAM tends to the constant $e^{L_1 r} - 1$ at large gradients and degenerates as $\|\nabla f(x)\| \to 0$. The bound for USAM tends to ρL_0 as $\|\nabla f(x)\| \to 0$ and grows exponentially as $\|\nabla f(x)\| \to +\infty$. This suggests that SAM's distortion can explode when the gradient vanishes, while USAM's distortion can explode when the gradient grows. The following proposition makes the two failures precise.

Proposition 2.1. Both perturbations can fail on one-dimensional convex objectives: SAM admits a nonzero two-cycle on a quadratic, and USAM diverges for large initializations on a quartic.

(i) Let $f(x) = x^2/2$, which satisfies (2) with $L = 1$. Run SAM with $r > 0$ and $\eta \in (0, 2)$. For every $x \neq 0$,

$$D_{\mathrm{SAM}}(x) := D(x, r\,\mathrm{sign}(x)) = \frac{r}{|x|}, \qquad x_{t+1} = \left[1 - \eta\left(1 + D_{\mathrm{SAM}}(x_t)\right)\right] x_t. \tag{8}$$

Let $a = \eta r/(2-\eta)$. Then $|x_{t+1}| < |x_t|$ if $|x_t| > a$, $|x_{t+1}| > |x_t|$ if $0 < |x_t| < a$, and $x_{t+1} = -x_t$ if $|x_t| = a$. Consequently $\pm a$ is a two-cycle, and no trajectory converges to the minimizer (unless it lands on it exactly).

(ii) Let $f(x) = x^4/4$, which satisfies (3) with $(L_0, L_1) = (4, 1)$. Run USAM with $\rho > 0$ and $\eta > 0$. For every $x \neq 0$,

$$D_{\mathrm{USAM}}(x) := D(x, \rho x^3) = (1+\rho x^2)^3 - 1, \qquad x_{t+1} = \left[1 - \eta\left(1 + D_{\mathrm{USAM}}(x_t)\right) x_t^2\right] x_t. \tag{9}$$

Consequently, if $|x_0| > (2/(\eta\rho^3))^{1/8}$, then x_{t+1} has the opposite sign of x_t and $|x_{t+1}| > |x_t|$ for every t, and $|x_t| \to \infty$.

The quadratic construction can be viewed as an exact illustration of established late-stage phenomena: Bartlett et al. (2023) showed that fixed-radius SAM typically approaches a two-cycle on convex quadratics, and Si & Yun (2023) established unavoidable stationarity neighborhoods for practical fixed-radius SAM in several settings. Our point is the complementarity: the distortion can become unbounded at opposite ends of the gradient scale, with nonconvergence for these fixed-parameter updates. The following theorem addresses all perturbations along the gradient whose nonnegative multiplier depends only on the gradient norm: it gives a necessary condition for uniform boundedness of the relative distortion over the whole class of objectives in (3).

Theorem 2.2. Let $L_0 > 0$, $L_1 > 0$, and consider the perturbation given by a nonnegative collinear rule $d(x) = \alpha(\|\nabla f(x)\|)\nabla f(x)$. Suppose that for every one-dimensional objective satisfying (3) and every point with nonzero gradient, $D(x, d(x)) \le \delta$. Then, for every $s > 0$,

$$\alpha(s) \le \min\left\{\frac{\delta}{L_0}, \frac{\log(1+\delta)}{L_1 s}\right\}. \tag{10}$$

Note that the theorem covers SAM with $\alpha(s) = r/s$ and USAM with $\alpha(s) = \rho$. The bound says that the perturbation length $\alpha(s)s$ must be at most linear in the gradient near $s = 0$ and at most a constant for large s. SAM violates the first requirement since $r/s \to \infty$ as $s \to 0$, and USAM violates the second since $\rho s \to \infty$ as $s \to \infty$. In words, every perturbation of this form with uniformly bounded distortion must be USAM-like near stationarity and SAM-like at large gradients. Dividing $\rho\nabla f(x)$

by $L_0 + L_1 \|\nabla f(x)\|$ gives the multiplier $\alpha(s) = \rho/(L_0 + L_1 s)$, which has both properties and lies below (10) for every s once $\delta = e^\rho - 1$. Satisfying the necessary condition alone does not establish bounded distortion. We prove that this construction is sufficient next, and extend it to the loss-dependent scale.

2.2 Admissible scales and the CS-SAM update

Theorem 2.2 motivates a perturbation that divides the gradient by a local curvature upper bound. In this subsection we isolate the properties of such a bound that our analysis uses, state the update, and prove that it keeps the distortion bounded for any scale with these properties.

Definition 2.3 (Admissible curvature scale)**.** Let $\kappa \geq 0$. A function $M : \mathbb{R}^d \to (0, \infty)$ is *κ-admissible* for f if, for all $x, y \in \mathbb{R}^d$,

$$\left\|\nabla^2 f(x)\right\| \leq M(x), \qquad M(y) \leq M(x)\, e^{\kappa\|y-x\|}, \qquad \kappa\,\|\nabla f(x)\| \leq M(x). \tag{11}$$

Let us read the three conditions in (11). The first says that M is a curvature scale, that is, it upper-bounds the local curvature everywhere. The second says that the scale varies at most exponentially in the distance travelled, so that a short move changes it by a bounded factor. The third ties the scale to the gradient norm; it is what makes a step proportional to $\nabla f(x)/M(x)$ short, so that it stays under control along the perturbation. Under (3), the scale $M_L(x) = L_0 + L_1 \|\nabla f(x)\|$ is L_1-admissible, so the concrete scale suggested by Theorem 2.2 is an instance of this definition. The following proposition records this together with the corresponding facts for the scales of (2) and (4).

Proposition 2.4 (Properties of the curvature scales)**.**

(i) Under (2), the constant scale $M \equiv L$ is admissible with $\kappa = 0$.

(ii) Under (3), the scale M_L is admissible with $\kappa = L_1$.

(iii) Under (4), either $M_H \equiv 0$ and f is constant, or $M_H > 0$ everywhere and the scale M_H is admissible with $\kappa = \sqrt{H_1}$.

(iv) Under (3), for every $x \in \mathbb{R}^d$,

$$L_0 + L_1 \|\nabla f(x)\| \leq \frac{3}{2} L_0 + \frac{3L_1^2}{2\,(2\log 2 - 1)} \big(f(x) - f_\star\big). \tag{12}$$

In particular, (3) implies (4) with the constants $H_0 = 3L_0/2$ and $H_1 = 3L_1^2/(2(2\log 2 - 1))$.

Items (i)–(iii) say that the three curvature bounds in (2), (3) and (4) are admissible, with $\kappa = 0$, $\kappa = L_1$, and $\kappa = \sqrt{H_1}$ respectively; the only exception is the trivial case of a constant f, for which M_H vanishes and every point is a global minimizer. Item (iv) is an inclusion between the two general models: every objective satisfying (3) also satisfies (4), with $H_0 = 3L_0/2$ and $H_1 = 3L_1^2/(2(2\log 2 - 1))$. This inclusion also appears as Proposition B.1 of Alimisis et al. (2026) and as Proposition 3 of Vaswani & Babanezhad Harikandeh (2025), with different constants.

The update. Given an admissible curvature scale M, parameters $\rho \geq 0$ and $\gamma > 0$, and a starting point x_0, we propose Curvature-Scaled SAM, which iterates

$$y_t = x_t + \frac{\rho}{M(x_t)} \nabla f(x_t), \qquad x_{t+1} = x_t - \frac{\gamma}{M(y_t)} \nabla f(y_t). \tag{CS-SAM}$$

Here, $\rho \geq 0$ sets the perturbation and $\gamma > 0$ the descent step. The perturbation step $\rho/M(x_t)$ and the descent step $\gamma/M(y_t)$ use the *same* curvature scale, each evaluated at the point where the corresponding gradient is taken. If $\nabla f(x_t) = 0$ the method stops at x_t. Two instances are of particular interest and are compared in Table 2. With the gradient scale $M = M_L$, the perturbation step is $\rho/(L_0 + L_1 \|\nabla f(x_t)\|)$ and the descent step is $\gamma/(L_0 + L_1 \|\nabla f(y_t)\|)$; in particular, the perturbation multiplier is exactly the $\alpha(s) = \rho/(L_0 + L_1 s)$ permitted by Theorem 2.2. With the loss scale $M = M_H$, the perturbation step is $\rho/(H_0 + H_1(f(x_t) - f_\star))$ and the descent step is $\gamma/(H_0 + H_1(f(y_t) - f_\star))$. Note that in this case, the perturbation multiplier is *not* covered by Theorem 2.2. Like SAM and USAM, CS-SAM costs two gradient evaluations per iteration; the scale M_L is computed from the gradient already in hand, and M_H from the function value and $f_\star$.

Table 2: The curvature scales associated with (2), (3) and (4) as instances of CS-SAM. For each scale we list the assumption under which it is a curvature scale, the admissibility constant κ of Proposition 2.4, the perturbation and descent steps of CS-SAM, and the curvature-scaled gradient method obtained at $\rho = 0$ (CS-GD, Section 2.2).

Curvature scale $M(x)$	**Assumption**	κ	**Perturbation step** $\rho/M(x_t)$	**Descent step** $\gamma/M(y_t)$	**CS-SAM at** $\rho = 0$ **(CS-GD)**
L (constant)	(2)	0	ρ/L	γ/L	gradient descent with step γ/L
$M_L(x) = L_0 + L_1 \|\nabla f(x)\|$	(3)	L_1	$\dfrac{\rho}{L_0 + L_1 \|\nabla f(x_t)\|}$	$\dfrac{\gamma}{L_0 + L_1 \|\nabla f(y_t)\|}$	the gradient method of Vankov et al. (2025), Gorbunov et al. (2025)
$M_H(x) = H_0 + H_1(f(x) - f_\star)$	(4)	$\sqrt{H_1}$	$\dfrac{\rho}{H_0 + H_1(f(x_t) - f_\star)}$	$\dfrac{\gamma}{H_0 + H_1(f(y_t) - f_\star)}$	a loss-dependent step similar to Alimisis et al. (2026)

Let us highlight the degenerate case. For $\rho = 0$ we have $y_t = x_t$, and CS-SAM reduces to curvature-scaled gradient descent, that is, gradient descent with the curvature-scaled step $\gamma/M(x_t)$:

$$x_{t+1} = x_t - \frac{\gamma}{M(x_t)} \nabla f(x_t). \tag{CS-GD}$$

CS-SAM is thus a generalization of CS-GD, whose instances at the three scales are listed in the last column of Table 2, and in Section 3 we compare its guarantees with known gradient descent results.

In contrast to SAM and USAM, the relative distortion of CS-SAM is uniformly bounded for *any* admissible curvature scale, as we now show.

Theorem 2.5. Let M be κ-admissible, let $\rho \geq 0$, $\gamma > 0$, and run CS-SAM. Then every iteration with $\nabla f(x_t) \neq 0$ satisfies

$$e^{-\rho} M(x_t) \leq M(y_t) \leq e^{\rho} M(x_t), \qquad \frac{\|\nabla f(y_t) - \nabla f(x_t)\|}{\|\nabla f(x_t)\|} \leq e^{\rho} - 1. \tag{13}$$

In particular, $D_{\text{CS-SAM}} \leq e^{\rho} - 1$.

Let us highlight what Theorem 2.5 says. The curvature scale at the perturbed point stays within a factor $e^{\pm\rho}$ of the scale at the iterate, and the relative gradient distortion never exceeds $e^{\rho} - 1$. This is exactly the uniform control that (6), (7), and Proposition 2.1 show SAM and USAM lack: the bound does not degenerate as $\|\nabla f(x_t)\| \to 0$, nor as $\|\nabla f(x_t)\| \to \infty$. Moreover, the bound depends neither on the admissibility constant κ nor on which admissible scale is used, so the constant scale, the (L_0, L_1) scale, and the (H_0, H_1) scale all inherit it. In particular, taking $\delta = e^{\rho} - 1$ in Section 2.1, $D_{\text{CS-SAM}} < 1$ whenever $\rho < \log 2$, which is a necessary condition for the rates of Section 3: the condition $\chi(\rho, \gamma) > 0$ of (14) under which they hold forces $\rho < \log 2$.

3 Convergence Rates

In this section, we present convergence rates for CS-SAM run with the curvature scales M_L of (3) and M_H of (4). We first fix the constants and the function classes that appear in the bounds. Then, Section 3.1 states the rates for the gradient scale M_L and Section 3.2 those for the loss scale M_H, each as one theorem with a nonconvex, a convex, and a PL item. The proofs can be found in Appendices D.1 and D.2; the auxiliary lemmas they rely on are collected in Appendix B.

Constants. For $u \geq 0$ let $\omega(u) := (e^u - 1 - u)/u^2$ for $u > 0$ and $\omega(0) := 1/2$. Define

$$\chi(\rho, \gamma) := 2 - e^{\rho} - \gamma\, \omega(\gamma)\, e^{3\rho}, \qquad c_0 := 1 - \frac{\rho\,(e^{\rho} - 1)}{2\log 2 - 1} - \frac{\gamma e^{3\rho}}{2\,(2\log 2 - 1)}. \tag{14}$$

The condition $\chi(\rho, \gamma) > 0$ forces $\rho < \log 2$. All results below assume $\chi(\rho, \gamma) > 0$; the convex results assume in addition $c_0 \geq 0$. The two conditions restrict how large the perturbation and the step may be taken. In the degenerate case $\rho = 0$, that is, for CS-GD, the constants reduce to $\gamma\chi(0, \gamma) = 1 + 2\gamma - e^{\gamma}$ and $c_0 = 1 - \gamma/(2(2\log 2 - 1))$.

Function classes. We say that f is *convex with a minimizer* if f is convex and attains $f_\star$ at some $x_\star$. We say that f satisfies the *μ-PL condition* for some $\mu > 0$ if

$$\tfrac{1}{2} \|\nabla f(x)\|^2 \geq \mu\big(f(x) - f_\star\big) \qquad \forall x \in \mathbb{R}^d. \tag{15}$$

3.1 (L_0, L_1) CURVATURE SCALE: $M = M_L$

Throughout this subsection f satisfies (3) and CS-SAM is run with the curvature scale $M = M_L$, so that the perturbation step is $\rho/(L_0 + L_1 \|\nabla f(x_t)\|)$ and the descent step is $\gamma/(L_0 + L_1 \|\nabla f(y_t)\|)$. We collect the three rates in one statement.

Theorem 3.1. Let f satisfy (3), let $\rho \geq 0$ and $\gamma > 0$ satisfy $\chi(\rho, \gamma) > 0$, and let $(x_t)_{t\geq 0}$ be generated by CS-SAM with $M = M_L$. Then, for every $T \geq 1$, the following hold:

(i) *(Nonconvex rate)*

$$\min_{0\leq t<T} \|\nabla f(x_t)\|^2 \leq \left(\sqrt{\frac{e^\rho L_0 \left(f(x_0) - f_\star\right)}{\gamma\chi(\rho,\gamma)\, T}} + \frac{e^\rho L_1 \left(f(x_0) - f_\star\right)}{\gamma\chi(\rho,\gamma)\, T} \right)^2. \tag{16}$$

(ii) *(Convex rate)* If in addition f is convex with a minimizer $x_\star$ and $c_0 \geq 0$, then

$$f(x_T) - f_\star \leq \frac{e^\rho \Big(L_0 \|x_0 - x_\star\|^2 + L_1 \|x_0 - x_\star\| \left(f(x_0) - f_\star\right) \Big)}{\gamma\chi(\rho,\gamma)\, T}. \tag{17}$$

(iii) *(PL rate)* If in addition f satisfies the μ-PL condition (15), then

$$f(x_T) - f_\star \leq \left(1 - \frac{2\mu\gamma e^{-\rho}\chi(\rho,\gamma)}{L_0 + L_1\sqrt{2\mu\left(f(x_0) - f_\star\right)}} \right)^T \left(f(x_0) - f_\star\right). \tag{18}$$

Let us read the three items in turn. Item (i) shows that the best squared gradient norm decays as $\mathcal{O}(1/T)$: expanding the square, the leading term is $e^\rho L_0(f(x_0) - f_\star)/(\gamma\chi(\rho,\gamma)T)$ and depends on L_0 alone, while the gradient-dependent constant L_1 enters only through terms of lower order in T. Note that the bound requires nothing beyond (3) and $\chi(\rho, \gamma) > 0$, that is, no bounded gradients and no bounded iterates, and that running the method needs nothing beyond the constants L_0, L_1 that define the scale. Item (ii) gives an $\mathcal{O}(1/T)$ bound on the gap of the last iterate for convex objectives, with a leading term proportional to $L_0 \|x_0 - x_\star\|^2$ and a second term proportional to $L_1 \|x_0 - x_\star\| (f(x_0) - f_\star)$. The additional condition $c_0 \geq 0$ is what keeps the iterates within the initial distance of $x_\star$ (Lemma B.6), and it is the only place where the second constant of (14) enters. Item (iii) gives a linear rate under the PL condition.

Comparison with gradient methods at $\rho = 0$. Let us now compare our guarantees with the known ones in the degenerate case. For $\rho = 0$, CS-SAM with $M = M_L$ is the gradient descent with step $\gamma/(L_0 + L_1 \|\nabla f(x_t)\|)$, and the constant $e^\rho/(\gamma\chi(\rho, \gamma))$ in items (i) and (ii) equals $1/(1 + 2\gamma - e^\gamma)$, which is $1/(2\log 2 - 1) \approx 2.59$ at $\gamma = \log 2$. This is the method analyzed by Vankov et al. (2025) and Gorbunov et al. (2025), and Theorem 3.1(i)–(ii) give the same rates as Theorems 3.1 and 3.2 of Vankov et al. (2025): the best squared gradient norm decays as $1/T$ with the L_1 term of lower order, and the gap decays as $1/T$ with an $L_1 \|x_0 - x_\star\|$ term. Our analysis therefore recovers the known asymptotic rate orders in T of the curvature-scaled gradient method as the special case $\rho = 0$, and extends them to the SAM setting. To the best of our knowledge, Theorem 3.1 provides the first convergence rates for a SAM-type method under the (L_0, L_1) curvature scale.

3.2 (H_0, H_1) CURVATURE SCALE: $M = M_H$

Throughout this subsection f satisfies (4) and is nonconstant, so that $M_H > 0$ everywhere by Proposition 2.4(iii), and CS-SAM is run with the curvature scale $M = M_H$: the perturbation step is $\rho/(H_0 + H_1(f(x_t) - f_\star))$ and the descent step is $\gamma/(H_0 + H_1(f(y_t) - f_\star))$. Note that the loss scale is evaluated relative to $f_\star$, which the method therefore needs, in contrast to the gradient scale of Section 3.1.

Theorem 3.2. Let f satisfy (4) and be nonconstant, let $\rho \geq 0$ and $\gamma > 0$ satisfy $\chi(\rho, \gamma) > 0$, and let $(x_t)_{t\geq 0}$ be generated by CS-SAM with $M = M_H$. Then, for every $T \geq 1$, the following hold:

(i) *(Nonconvex rate)*

$$\min_{0\le t<T} \|\nabla f(x_t)\|^2 \le \frac{e^{\rho}\big(H_0 + H_1(f(x_0)-f_\star)\big)\big(f(x_0)-f_\star\big)}{\gamma\chi(\rho,\gamma)\,T}. \tag{19}$$

(ii) *(Convex rate)* If in addition f is convex with a minimizer $x_\star$ and $c_0 \ge 0$, then

$$f(x_T) - f_\star \le \frac{e^{\rho}\big(H_0 + H_1(f(x_0)-f_\star)\big)\,\|x_0 - x_\star\|^2}{\gamma\chi(\rho,\gamma)\,T}. \tag{20}$$

(iii) *(PL rate)* If in addition f satisfies the μ-PL condition (15), then

$$f(x_T) - f_\star \le \left(1 - \frac{2\mu\gamma e^{-\rho}\chi(\rho,\gamma)}{H_0 + H_1\big(f(x_0)-f_\star\big)}\right)^T \big(f(x_0)-f_\star\big). \tag{21}$$

The three items have the same shape as in Theorem 3.1, with one difference worth pointing out: the dependence on the curvature parameters is different. Item (i) bounds the best squared gradient norm by $\mathcal{O}(1/T)$ with a constant proportional to $(H_0+H_1(f(x_0)-f_\star))(f(x_0)-f_\star)$. Thus, unlike the L_1 contribution in Theorem 3.1(i), the H_1 contribution occurs at the leading $1/T$ order. Item (ii) gives an $\mathcal{O}(1/T)$ bound on the last-iterate gap for convex objectives, proportional to $(H_0 + H_1(f(x_0) - f_\star))\,\|x_0 - x_\star\|^2$, again under $c_0 \ge 0$, which bounds the iterates as in Section 3.1. Item (iii) gives a linear rate under the PL condition, with contraction factor $1-2\mu\gamma e^{-\rho}\chi(\rho,\gamma)/(H_0+H_1(f(x_0)-f_\star))$.

Comparison with gradient methods at $\rho = 0$. For $\rho = 0$, CS-SAM with $M = M_H$ is gradient descent with the step $\gamma/(H_0 + H_1(f(x_t) - f_\star))$, and, as in Section 3.1, the constant $e^{\rho}/(\gamma\chi(\rho,\gamma))$ in items (i) and (ii) equals $1/(1 + 2\gamma - e^{\gamma})$, which is $1/(2\log 2 - 1) \approx 2.59$ at $\gamma = \log 2$. This kind of loss-dependent step was used by Alimisis et al. (2026), who took $1/(10H_0 + 20H_1(f(x_t) - f_\star))$, and Theorem 3.2(ii)–(iii) give the same rates as their Theorems 4.2 and 4.3, namely a $1/T$ gap in the convex case and a linear rate under the PL condition, with a larger step at $\gamma = \log 2$: their denominator coefficients 10 and 20 are both replaced by $1/\log 2 \approx 1.44$. At $\rho = 0$ our guarantees thus match the known rates for gradient descent with a loss-dependent step and allow larger step sizes, and they extend to any positive perturbation with $\chi(\rho,\gamma) > 0$. To the best of our knowledge, Theorem 3.2 provides the first convergence rates for a SAM-type method under the (H_0, H_1) curvature scale.

Remark 3.3. As we show in Proposition 2.4(iv), every objective satisfying (3) also satisfies (4) with $H_0 = 3L_0/2$ and $H_1 = 3L_1^2/(2(2\log 2 - 1))$, so one could ask why we treat the two scales separately. Note that CS-SAM with $M = M_L$ and CS-SAM with $M = M_H$ are the same algorithm run with different steps: while one assumption implies the other, the perturbation step and the descent step are scaled by M_L in one case and by M_H in the other. Theorems 3.1 and 3.2 are therefore guarantees for two different step choices.

4 Numerical Experiments

In this section, we first verify the results of Section 2 on the two one-dimensional objectives of Proposition 2.1 (Section 4.1), and we then test CS-SAM on the training of deep neural networks against tuned SAM and USAM baselines (Section 4.2). The complete protocols are given in Appendix E.

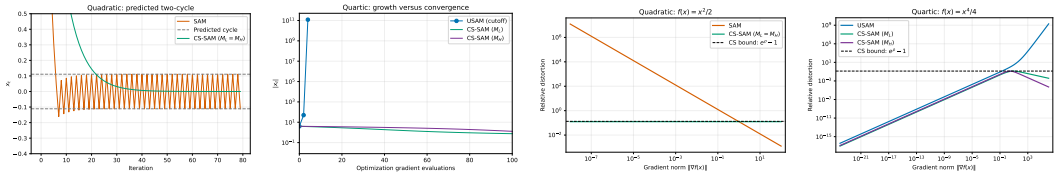
4.1 Verification of Section 2

The experiments of this subsection are deterministic and one-dimensional, with exact function values and gradients. We use the quadratic $f(x) = x^2/2$ and the quartic $f(x) = x^4/4$ of Proposition 2.1. The two scales coincide on the quadratic, $M_L = M_H \equiv 1$, while on the quartic $M_L(x) = 4 + |x|^3$ and $M_H(x) = 3 + 3x^4/4$, that is, the constants $(L_0, L_1) = (4, 1)$ of Proposition 2.1(ii) and $(H_0, H_1) = (3, 3)$; both bound the curvature $3x^2$ at every x (Appendix E.1). Throughout, CS-SAM is run with $\rho = \gamma = 1/8$, for which $\chi(\rho,\gamma) > 0$ and $c_0 \ge 0$. The results are collected in Figure 1.

Two-cycle and divergence. We first run the three methods on the two objectives. On the quadratic, SAM with $r = 1$ and $\eta = 0.2$ is started at $x_0 = 3$; by Proposition 2.1(i), its iterates approach the two-cycle $\pm a$ with $a = \eta r/(2 - \eta) = 1/9$. On the quartic, USAM with $\rho = 1/8$ and $\eta = 1/32$ is started at $x_0 = 4$, which exceeds the threshold $(2/(\eta\rho^3))^{1/8} \approx 3.67$ of Proposition 2.1(ii). Figures 1a

and 1b show the iterates. We observe that SAM oscillates between $\pm 1/9$ (dashed lines) and does not converge, and that USAM changes sign and grows at every step, reaching $|x_2| \approx 10^{11}$ after two updates, at which point we stop the run. In contrast, CS-SAM converges to the minimizer on both objectives, with either scale.

Uniform distortion control. We next evaluate the relative gradient distortion (5) of each perturbation as a function of the gradient norm, no outer step enters this measurement. Here SAM uses $r = 1/8$ on the quadratic and USAM uses $\rho = 1/8$ on the quartic, so that all three perturbations share the numerical parameter $1/8$. Figures 1c and 1d show the distortion together with the bound $e^{\rho} - 1 \approx 0.133$ of Theorem 2.5. We observe the two failures of Section 2.1: the distortion of SAM equals $r/\|\nabla f(x)\|$ and grows without bound as the gradient vanishes, while that of USAM equals $(1 + \rho x^2)^3 - 1$ and grows without bound as the gradient grows. The distortion of CS-SAM stays below $e^{\rho} - 1$ at every gradient norm and for both scales, as Theorem 2.5 guarantees. Note that on the quadratic it equals ρ exactly, and on the quartic it peaks below the bound and decays at both ends.



(a) Quadratic: iterates (b) Quartic: $|x_t|$ (c) Quadratic: distortion (d) Quartic: distortion

Figure 1: Verification of Section 2 on the objectives of Proposition 2.1. (a) Iterates of SAM on $f(x) = x^2/2$ against the predicted two-cycle (dashed), with CS-SAM. (b) $|x_t|$ of USAM on $f(x) = x^4/4$ against the number of gradient evaluations, with CS-SAM for both scales. (c)–(d) Relative gradient distortion (5) against the gradient norm and the bound $e^{\rho} - 1$ of Theorem 2.5.

4.2 DNN EXPERIMENTS

In this subsection, we test CS-SAM on the training of deep neural networks and compare it with tuned SAM and USAM baselines. We go beyond the deterministic setting of Section 4.1 and of our guarantees: the networks are trained with minibatch gradients, and the curvature scales are evaluated on minibatch quantities. The experiments are designed to test whether the two scales are competitive in practice. The complete grids and the per-run configurations are given in Appendix E.2.

We train ResNet-20 and ResNet-32 with batch normalization on CIFAR-10 and CIFAR-100 for 200 epochs with batch size 128, random-crop and horizontal-flip augmentation without cutout, and the cross-entropy loss. Momentum and weight decay are 0 for every method, so that the comparison isolates the update rules. In the minibatch setting the M_L *scale* is $L_0 + L_1 \|\nabla f_i(x)\|$, and the M_H *scale* is $H_0 + H_1[f_i(x) - \ell_i^\star]_+$, where $\ell_i^\star$ is a lower bound of f_i and $[z]_+ = \max\{z, 0\}$. We set $\ell_i^\star = 0$ in all our expriments. We compare against SAM and USAM, each with a constant and with a cosine annealing schedule, tuned on a 4×4 grid of learning rate $\times$ radius (respectively multiplier) for every model–dataset cell. For CS-SAM we tune the M_L scale over $\rho \in \{0.2, 0.4, 0.8, 1.6\}$, $\gamma \in \{0.2, 0.4, 0.8\}$, $L_0 \in \{0.02, 0.5, 2\}$, and the M_H scale over $\rho, \gamma \in \{0.3, 1, 3\}$, $H_1 \in \{1, 3, 10, 30\}$. We run each experiment for 200 epochs. Table 3 reports, for each method and cell, the mean and standard deviation over the three seeds of the best test accuracy. We observe that both curvature scales are competitive: CS-SAM with M_H achieves the highest mean best-test accuracy in all four settings, while CS-SAM with M_L beats tuned SAM and USAM in three out of four settings.

Table 3: Best test accuracy (%, maximum over 200 epochs), mean $\pm$ standard deviation over 3 seeds, for ResNet-20/32 trained on CIFAR-10/100. Bold marks the best mean in each column.

Method	ResNet-20 / CIFAR-10	ResNet-32 / CIFAR-10	ResNet-20 / CIFAR-100	ResNet-32 / CIFAR-100
SAM, constant lr	$90.60_{\pm 0.20}$	$91.24_{\pm 0.08}$	$64.66_{\pm 0.61}$	$65.55_{\pm 0.62}$
SAM, cosine	$90.49_{\pm 0.16}$	$91.16_{\pm 0.23}$	$65.32_{\pm 0.21}$	$66.07_{\pm 0.69}$
USAM, constant lr	$90.15_{\pm 0.30}$	$90.91_{\pm 0.11}$	$64.73_{\pm 0.12}$	$65.61_{\pm 0.40}$
USAM, cosine	$90.13_{\pm 0.41}$	$90.84_{\pm 0.19}$	$65.33_{\pm 0.13}$	$66.77_{\pm 0.34}$
CS-SAM with M_L	$90.87_{\pm 0.30}$	$91.72_{\pm 0.33}$	$65.18_{\pm 0.42}$	$66.80_{\pm 0.59}$
CS-SAM with M_H	$\mathbf{91.01}_{\pm \mathbf{0.06}}$	$\mathbf{91.73}_{\pm \mathbf{0.22}}$	$\mathbf{65.68}_{\pm \mathbf{0.34}}$	$\mathbf{67.06}_{\pm \mathbf{0.44}}$

AI USE STATEMENT

We used generative AI tools to suggest proof ideas, which the authors checked and wrote up, and to streamline and edit the wording of some paragraphs and sentences. We take responsibility for the final content of this work.

REFERENCES

Samir Adly and Ba Khiet Le. Stationarity floors and vanishing perturbations in sharpness-aware minimization. *arXiv preprint arXiv:2608.06692*, 2026.

Foivos Alimisis, Rustem Islamov, and Aurelien Lucchi. Why do we need warm-up? a theoretical perspective. In *ICML*, 2026.

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *ICML*, 2022.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In *ICML*, 2023.

Peter L. Bartlett, Philip M. Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *Journal of Machine Learning Research*, 24(316):1–36, 2023.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *ICML*, 2023.

Yifei Cheng, Li Shen, Hao Sun, Nan Yin, Xiaochun Cao, and Enhong Chen. LightSAM: Parameter-agnostic sharpness-aware minimization. *arXiv preprint arXiv:2505.24399*, 2025.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signSGD. In *NeurIPS*, 2022.

Yan Dai, Kwangjun Ahn, and Suvrit Sra. The crucial role of normalization in sharpness-aware minimization. In *NeurIPS*, 2023.

Jiaxin Deng and Junbiao Pang. On the implicit flatness bias of sharpness-aware minimization: A linear stability analysis with quantitative hyperparameter bounds. *arXiv preprint arXiv:2608.03197*, 2026.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. In *ICLR*, 2025.

T. H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919.

Moritz Haas, Jin Xu, Volkan Cevher, and Leena Chennuru Vankadara. μP^2: Effective sharpness aware minimization requires layerwise perturbation scaling. In *NeurIPS*, 2024.

Zhen Huang, Jiaxin Deng, and Junbiao Pang. Gradient-energy guided block-wise perturbations for sharpness-aware minimization. *arXiv preprint arXiv:2607.18306*, 2026.

Pham Duy Khanh, Hoang-Chau Luong, Boris S. Mordukhovich, and Dat Ba Tran. Fundamental convergence analysis of sharpness-aware minimization. In *NeurIPS*, 2024.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In *ICML*, 2022.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *ICML*, 2023.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *ICML*, 2021.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *NeurIPS*, 2023.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Dimitris Oikonomou and Nicolas Loizou. Sharpness-aware minimization: General analysis and improved rates. In *ICLR*, 2025.

Dimitris Oikonomou and Nicolas Loizou. Adaptive sharpness-aware minimization with a polyak-type step size: A theory-grounded scheduler. In *ICML*, 2026.

Chanwoong Park, Uijeong Jang, Ernest K. Ryu, and Insoon Yang. Sharpness-aware minimization can hallucinate minimizers. In *ICML*, 2026.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Andrei Semenov, Martin Jaggi, and Nikita Doikov. Gradient-normalized smoothness for optimization with approximate hessians. In *ICLR*, 2026.

Dongkuk Si and Chulhee Yun. Practical sharpness-aware minimization cannot converge all the way to optima. In *NeurIPS*, 2023.

Sidak Pal Singh, Hossein Mobahi, Atish Agarwala, and Yann Dauphin. Avoiding spurious sharpness minimization broadens applicability of SAM. In *ICML*, 2025.

Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: A recipe for newton-type methods. *Mathematical Programming*, 178(1–2):145–213, 2019.

Hossein Taheri and Christos Thrampoulidis. Fast convergence in learning two-layer neural networks with separable data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9944–9952, 2023.

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U. Stich. Optimizing (l_0, l_1)-smooth functions by gradient methods. In *ICLR*, 2025.

Sharan Vaswani and Reza Babanezhad Harikandeh. Armijo line-search can make (stochastic) gradient descent provably faster. In *ICML*, 2025.

Sharan Vaswani, Yifan Sun, and Reza Babanezhad. Convergence of steepest descent and Adam under non-uniform smoothness. In *ICML*, 2026.

Jinping Wang, Qinhan Liu, Zhiwu Xie, and Zhiqiang Gao. Fix the loss, not the radius: Rethinking the adversarial perturbation of sharpness-aware minimization. In *ICML*, 2026.

Xuehao Wang, Weisen Jiang, Shuai Fu, and Yu Zhang. Enhancing sharpness-aware minimization by learning perturbation radius. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2024.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. In *NeurIPS*, 2020a.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *ICLR*, 2020b.

Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. In *ICLR*, 2025.

Supplementary Material

The appendix is organized as follows. In Appendix A, we give a more detailed discussion of related work. In Appendix B, we collect the auxiliary lemmas on admissible curvature scales and on a single iteration of CS-SAM that all proofs rely on. Appendix C presents the proofs for Section 2: the verification of the distortion bounds (6) and (7), and the proofs of Proposition 2.1, Theorem 2.2, Proposition 2.4, and Theorem 2.5. Appendix D presents the proofs for Section 3, with Appendix D.1 for the (L_0, L_1) curvature scale (Theorem 3.1) and Appendix D.2 for the (H_0, H_1) curvature scale (Theorem 3.2). Finally, Appendix E gives the experimental details for Section 4.

CONTENTS

1 Introduction **1**

1.1 Main Contributions 2

2 From gradient distortion to curvature-scaled SAM **3**

2.1 Relative gradient distortion and the failures of SAM and USAM 3

2.2 Admissible scales and the CS-SAM update 5

3 Convergence rates **6**

3.1 (L_0, L_1) curvature scale: $M = M_L$ 7

3.2 (H_0, H_1) curvature scale: $M = M_H$ 7

4 Numerical experiments **8**

4.1 Verification of Section 2 8

4.2 DNN experiments 9

A Further Related Work **13**

B Auxiliary lemmas **15**

C Proofs for Section 2 **19**

D Proofs for Section 3 **22**

D.1 Proofs for Section 3.1: the (L_0, L_1) curvature scale 22

D.2 Proofs for Section 3.2: the (H_0, H_1) curvature scale 23

E Experimental details **25**

E.1 Verification experiments of Section 4.1 25

E.2 Deep-network experiments of Section 4.2 26

A Further Related Work

In this section, we complement the discussion of related work in Section 1 with a more detailed survey. We proceed from the broader literature on the dynamics of SAM and on adaptive perturbation rules to the lines closest to our work: relative control of the perturbation-induced gradient error, optimization under the curvature models (3) and (4), and relative gradient models. We close each paragraph with a brief comparison with our results.

SAM, normalization, and optimization dynamics. Sharpness-Aware Minimization replaces the gradient at the current iterate with a gradient evaluated after a parameter perturbation, motivated by minimizing the loss over a fixed local neighborhood (Foret et al., 2021). Subsequent work has shown that its behavior cannot be understood solely as ordinary descent on a smooth objective. Andriushchenko & Flammarion (2022) studied the implicit bias of SAM and the role of stochastic sharpness, while Dai et al. (2023) identified stabilization and continued motion along manifolds of minima as important effects of the gradient normalization. Bartlett et al. (2023) showed that fixed-radius SAM typically approaches a two-cycle on convex quadratics and that analogous oscillations can induce drift toward regions with smaller Hessian spectral norm on nonquadratic landscapes. Si & Yun (2023) established unavoidable stationarity neighborhoods for practical fixed-radius SAM in several settings, and Park et al. (2026) showed that SAM can converge to points at which the perturbed-point gradient vanishes even though the original gradient does not. Our quadratic construction in Proposition 2.1(i) should therefore be viewed as an exact illustration of established late-stage phenomena. Our new point is that the low-gradient failure of SAM and the high-gradient amplification of USAM are complementary consequences of expressing the perturbation budget in inappropriate units, which Theorem 2.2 makes precise.

Unified and adaptive perturbation rules. Unified SAM (Oikonomou & Loizou, 2025) interpolates between SAM and USAM through a fixed normalization rule and provides deterministic and stochastic convergence guarantees under several sampling and step-size choices. Other methods adapt the perturbation: ASAM (Kwon et al., 2021) rescales the parameter-space neighborhood to obtain scale invariance, whereas Fisher SAM (Kim et al., 2022) replaces the Euclidean ball by a Fisher-information geometry. LETS (Wang et al., 2024) learns the perturbation radius through a bilevel training–validation objective, LightSAM (Cheng et al., 2025) adapts the perturbation radius and the outer learning rate using AdaGrad- and Adam-style statistics, and LE-SAM (Wang et al., 2026) replaces the fixed parameter-space radius by a fixed loss-space budget. More recent methods adapt the radius from the observed Taylor-approximation error (Deng & Pang, 2026) or redistribute a fixed global radius across parameter blocks using moving averages of blockwise gradient energy (Huang et al., 2026). Polyak-type SAM schedulers instead adapt the outer step from function values under classical smoothness (Oikonomou & Loizou, 2026). In contrast, CS-SAM adapts the perturbation through one scalar quantity, an upper bound on the local curvature (Definition 2.3), rather than through validation performance, optimizer statistics, parameter scale, blockwise gradient energy, or a loss-space budget. Moreover, the scaling of the perturbation is derived rather than chosen: it is motivated by the necessary condition of Theorem 2.2 for uniformly bounded relative gradient distortion, and Theorem 2.5 shows that it meets this requirement for every admissible scale. Finally, the same curvature scale sets the outer step, so that a single quantity controls both the perturbation and the descent step.

Relative-error control and exact stationarity. Inexact-gradient analysis provides a natural framework for studying gradients evaluated at perturbed points. Khanh et al. (2024) used this perspective to establish convergence results for several normalized and unnormalized SAM variants, including conditions that control the perturbation-induced gradient error. Adly & Le (2026) analyzed the deterministic family

$$y_t = x_t + \rho \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|^{\alpha}}, \qquad 0 \le \alpha \le 1,$$

under classical smoothness (2). They identified stationarity floors for constant-parameter rules and proposed a clipped perturbation that agrees with the original rule away from stationarity but becomes gradient-proportional near stationary points, yielding $\|\nabla f(x_t)\| \to 0$. Relative control of the gradient error is therefore not new by itself. Our contribution is to derive the scaling required to enforce such control uniformly under gradient-dependent curvature: by Theorem 2.2, every perturbation in the

gradient direction with uniformly bounded distortion over the class (3) must be USAM-like near stationarity and SAM-like at large gradients. The curvature-scaled perturbation of CS-SAM realizes both regimes and couples the resulting inner perturbation to a curvature-scaled outer step.

Optimization beyond uniform smoothness. The (L_0, L_1)-smoothness model (3) was introduced to describe objectives whose local curvature can grow with the gradient norm and to explain the effectiveness of clipping and normalized gradients (Zhang et al., 2020b). Subsequent work developed sharper clipping, acceleration, normalization, and adaptive methods for this class (Zhang et al., 2020a; Crawshaw et al., 2022; Chen et al., 2023; Koloskova et al., 2023; Li et al., 2023; Gorbunov et al., 2025; Vankov et al., 2025). Curvature-scaled ordinary gradient steps are therefore established machinery rather than a contribution of this paper. The loss-dependent condition (4) has likewise been used to explain learning-rate warm-up (Alimisis et al., 2026) and to analyze line search, steepest descent, and adaptive methods (Vaswani & Babanezhad Harikandeh, 2025; Vaswani et al., 2026). Our contribution is to use the same machinery for the perturbation: in CS-SAM, the curvature scale that sets the descent step also sets the perturbation, and Theorem 2.5 shows that this keeps the relative gradient distortion uniformly bounded. The resulting rates have the same asymptotic orders in T as those of curvature-scaled gradient descent, with different dependence on the curvature parameters and initial conditions, and they recover the known gradient-descent rates at $\rho = 0$ (Section 3).

Relative gradient models and curvature geometry. Gradient-normalized smoothness characterizes the neighborhood in which the gradient admits a reliable relative approximation, and uses it to globalize methods based on approximate Hessians (Semenov et al., 2026). We specialize this relative-locality principle to a first-order sharpness-aware perturbation, with an explicit rule based on a scalar curvature scale and a necessity result for perturbations in the gradient direction (Theorem 2.2). The exponential local models of (11) are also reminiscent of generalized self-concordance, where controlled curvature growth replaces a uniform Hessian bound (Sun & Tran-Dinh, 2019); unlike that convex, Newton-oriented framework, our analysis certifies a two-gradient SAM update on nonconvex objectives.

Geometry, generalization, and scope. Parameter-space sharpness is sensitive to neural-network reparameterization, and empirical evaluations show that no single sharpness measure universally predicts generalization (Kwon et al., 2021; Andriushchenko et al., 2023). Fisher SAM and μP^2 further illustrate that information-geometric or layerwise perturbation scaling can be important in modern networks (Kim et al., 2022; Haas et al., 2024). Moreover, late-stage SAM can be especially effective for selecting flatter minima (Zhou et al., 2025), while standard parameter-space SAM can optimize spurious notions of sharpness in language modeling (Singh et al., 2025). CS-SAM should therefore be viewed as an optimization method rather than a parameterization-invariant theory of generalization. The present analysis is deterministic, assumes exact gradients and function values, and uses a global Euclidean scalar curvature scale.

B AUXILIARY LEMMAS

In this section, we collect the lemmas used in Appendices C and D. All of them are stated for a general κ-admissible scale M (Definition 2.3), with three exceptions: Lemma B.1 is the differential form of Grönwall's inequality, Lemmas B.7 and B.8 concern the loss scale M_H of (4) and Lemma B.9 is an elementary fact about sequences.

Conventions for the proofs. The constants $\chi(\rho,\gamma)$ and c_0 are defined in (14). If the method stops at some x_t, then $\nabla f(x_t)=0$: the best gradient norm is then zero, and under convexity or the PL condition x_t is a global minimizer, so every bound below holds trivially from that index on. We therefore assume in all proofs that $\nabla f(x_t)\neq 0$ for every t. Note that $\chi(\rho,\gamma)>0$ implies $2-e^{\rho}>\gamma\omega(\gamma)e^{3\rho}>0$, hence $\rho<\log 2$ and $\gamma e^{-\rho}\chi(\rho,\gamma)>0$. Finally, recall that $\omega(u)=\sum_{k\geq 0}u^k/(k+2)!$, so ω is increasing on $[0,\infty)$, and that

$$u+u^2\omega(u)=e^u-1,\qquad u-u^2\omega(u)=1+2u-e^u\qquad (u\geq 0). \tag{22}$$

We begin with the differential form of Grönwall's inequality (Gronwall, 1919), stated for absolutely continuous functions because the gradient scale M_L is only Lipschitz along a segment. We include the short proof for completeness.

Lemma B.1 (Grönwall's inequality; see Gronwall, 1919)**.** Let $c\geq 0$ and let $g:[0,1]\to[0,\infty)$ be absolutely continuous with $g'(s)\leq c\,g(s)$ for almost every $s\in[0,1]$. Then $g(s)\leq g(0)e^{cs}$ for every $s\in[0,1]$.

Proof. The function $h(s):=e^{-cs}g(s)$ is absolutely continuous on $[0,1]$, with $h'(s)=e^{-cs}\big(g'(s)-c\,g(s)\big)\leq 0$ for almost every s. Hence $h(s)=h(0)+\int_0^s h'(u)\,du\leq h(0)=g(0)$, that is $g(s)\leq g(0)e^{cs}$. □

Lemma B.2. Let M be κ-admissible for f. Then, for all $x,d\in\mathbb{R}^d$,

$$\begin{aligned}\|\nabla f(x+d)-\nabla f(x)\|&\leq M(x)\,\frac{e^{\kappa\|d\|}-1}{\kappa\|d\|}\,\|d\|\,,\\ \big|f(x+d)-f(x)-\langle\nabla f(x),d\rangle\big|&\leq M(x)\,\omega(\kappa\|d\|)\,\|d\|^2\,,\end{aligned} \tag{23}$$

where the quotient is interpreted as 1 when $\kappa\|d\|=0$.

Proof. By the fundamental theorem of calculus and the first two admissibility conditions,

$$\begin{aligned}\|\nabla f(x+d)-\nabla f(x)\|&\leq\int_0^1\big\|\nabla^2 f(x+sd)\big\|\,\|d\|\,ds\\&\leq\|d\|\int_0^1 M(x)e^{\kappa s\|d\|}\,ds=M(x)\,\|d\|\,\frac{e^{\kappa\|d\|}-1}{\kappa\|d\|}.\end{aligned}$$

For the second inequality, write

$$f(x+d)-f(x)-\langle\nabla f(x),d\rangle=\int_0^1\langle\nabla f(x+sd)-\nabla f(x),d\rangle\,ds$$

and apply the first inequality with $d=sd$, which gives

$$\|\nabla f(x+sd)-\nabla f(x)\|\leq M(x)\,\|d\|\,\frac{e^{s\kappa\|d\|}-1}{\kappa\|d\|}.$$

Integrating,

$$\begin{aligned}\big|f(x+d)-f(x)-\langle\nabla f(x),d\rangle\big|&\leq M(x)\,\|d\|^2\int_0^1\frac{e^{s\kappa\|d\|}-1}{\kappa\|d\|}\,ds\\&=M(x)\,\|d\|^2\,\frac{e^{\kappa\|d\|}-1-\kappa\|d\|}{(\kappa\|d\|)^2}=M(x)\,\omega(\kappa\|d\|)\,\|d\|^2\,.\ \square\end{aligned}$$

Lemma B.3. Let M be κ-admissible for f. Then, for every $x \in \mathbb{R}^d$,

$$f(x) - f_\star \geq (2\log 2 - 1)\,\frac{\|\nabla f(x)\|^2}{M(x)}. \tag{24}$$

Proof. Fix $\alpha \geq 0$ and consider the point $x - \alpha\nabla f(x)/M(x)$. The third admissibility condition gives $\kappa\,\|\alpha\nabla f(x)/M(x)\| = \alpha\kappa\,\|\nabla f(x)\|/M(x) \leq \alpha$, so by Lemma B.2, the monotonicity of ω, and (22),

$$\begin{aligned} f_\star \leq f\Big(x - \alpha\frac{\nabla f(x)}{M(x)}\Big) &\leq f(x) - \alpha\frac{\|\nabla f(x)\|^2}{M(x)} + M(x)\,\omega(\alpha)\,\alpha^2\frac{\|\nabla f(x)\|^2}{M(x)^2} \\ &= f(x) - \big(\alpha - \alpha^2\omega(\alpha)\big)\frac{\|\nabla f(x)\|^2}{M(x)} = f(x) - \big(1 + 2\alpha - e^\alpha\big)\frac{\|\nabla f(x)\|^2}{M(x)}. \end{aligned}$$

Thus we get $f(x) - f_\star \geq (1 + 2\alpha - e^\alpha)\,\|\nabla f(x)\|^2/M(x)$ for every $\alpha \geq 0$. The coefficient is maximized at $\alpha = \log 2$, where it equals $2\log 2 - 1$. □

Lemma B.4. Let M be κ-admissible for f, let $\rho \geq 0$, and let x_t be an iterate of CS-SAM. Then

$$e^{-\rho}M(x_t) \leq M(y_t) \leq e^\rho M(x_t), \qquad \|\nabla f(y_t) - \nabla f(x_t)\| \leq (e^\rho - 1)\,\|\nabla f(x_t)\|, \tag{25}$$

$$\|\nabla f(y_t)\| \leq e^\rho\,\|\nabla f(x_t)\|, \qquad \langle \nabla f(x_t), \nabla f(y_t)\rangle \geq (2 - e^\rho)\,\|\nabla f(x_t)\|^2. \tag{26}$$

Proof. For $\rho = 0$ we have $y_t = x_t$ and all claims are trivial, so let $\rho > 0$. The third admissibility condition gives $\kappa\,\|y_t - x_t\| = \rho\kappa\,\|\nabla f(x_t)\|/M(x_t) \leq \rho$. The second condition, applied from x_t to y_t and from y_t to x_t, gives $M(y_t) \leq M(x_t)e^{\kappa\|y_t - x_t\|} \leq e^\rho M(x_t)$ and $M(x_t) \leq e^\rho M(y_t)$. Lemma B.2 with $d = y_t - x_t$ gives

$$\begin{aligned} \|\nabla f(y_t) - \nabla f(x_t)\| &\leq M(x_t)\,\frac{e^{\kappa\|y_t - x_t\|} - 1}{\kappa\,\|y_t - x_t\|}\,\frac{\rho\,\|\nabla f(x_t)\|}{M(x_t)} \\ &\leq \rho\,\|\nabla f(x_t)\|\,\frac{e^\rho - 1}{\rho} = (e^\rho - 1)\,\|\nabla f(x_t)\|, \end{aligned}$$

since $u \mapsto (e^u - 1)/u$ is increasing and $\kappa\,\|y_t - x_t\| \leq \rho$. The claims in (26) follow from the triangle and Cauchy–Schwarz inequalities:

$$\begin{aligned} \|\nabla f(y_t)\| &\leq \|\nabla f(x_t)\| + \|\nabla f(y_t) - \nabla f(x_t)\| \leq e^\rho\,\|\nabla f(x_t)\|, \\ \langle \nabla f(x_t), \nabla f(y_t)\rangle &= \|\nabla f(x_t)\|^2 + \langle \nabla f(x_t), \nabla f(y_t) - \nabla f(x_t)\rangle \\ &\geq \|\nabla f(x_t)\|^2 - (e^\rho - 1)\,\|\nabla f(x_t)\|^2. \end{aligned}$$

□

Lemma B.5. Let M be κ-admissible for f, let $\rho \geq 0$ and $\gamma > 0$ satisfy $\chi(\rho, \gamma) > 0$, and let x_t be an iterate of CS-SAM. Then

$$f(x_{t+1}) \leq f(x_t) - \chi(\rho,\gamma)\frac{\gamma}{M(y_t)}\,\|\nabla f(x_t)\|^2 \leq f(x_t) - \gamma e^{-\rho}\chi(\rho,\gamma)\,\frac{\|\nabla f(x_t)\|^2}{M(x_t)}. \tag{27}$$

In particular, $f(x_t) - f_\star$ is nonincreasing in t.

Proof. By the third admissibility condition at y_t, the displacement $x_{t+1} - x_t = -\gamma\nabla f(y_t)/M(y_t)$ satisfies

$$\kappa\,\|x_{t+1} - x_t\| = \gamma\kappa\,\frac{\|\nabla f(y_t)\|}{M(y_t)} \leq \gamma.$$

Lemma B.2 at x_t and the monotonicity of ω give

$$f(x_{t+1}) \leq f(x_t) - \frac{\gamma}{M(y_t)}\,\langle \nabla f(x_t), \nabla f(y_t)\rangle + M(x_t)\,\omega(\gamma)\,\frac{\gamma^2}{M(y_t)^2}\,\|\nabla f(y_t)\|^2.$$

By Lemma B.4, $\langle \nabla f(x_t), \nabla f(y_t)\rangle \geq (2-e^{\rho})\,\|\nabla f(x_t)\|^2$, $M(x_t)/M(y_t) \leq e^{\rho}$, and $\|\nabla f(y_t)\|^2 \leq e^{2\rho}\,\|\nabla f(x_t)\|^2$. Hence

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \frac{\gamma}{M(y_t)}\left[2 - e^{\rho} - \gamma\omega(\gamma)e^{3\rho}\right]\|\nabla f(x_t)\|^2 \\ &= f(x_t) - \chi(\rho,\gamma)\frac{\gamma}{M(y_t)}\,\|\nabla f(x_t)\|^2 . \end{aligned}$$

Since $\chi(\rho,\gamma) > 0$ and $M(y_t) \leq e^{\rho} M(x_t)$, the second inequality in (27) follows. □

Lemma B.6. Let f be convex with a minimizer $x_\star$, let M be κ-admissible for f, let $\rho \geq 0$ and $\gamma > 0$, and let x_t be an iterate of CS-SAM. Then

$$\|x_{t+1} - x_\star\|^2 \leq \|x_t - x_\star\|^2 - 2c_0\frac{\gamma}{M(y_t)}\big(f(x_t) - f_\star\big). \tag{28}$$

In particular, if $c_0 \geq 0$, then $\|x_t - x_\star\| \leq \|x_0 - x_\star\|$ for all t.

Proof. Expanding the square,

$$\|x_{t+1} - x_\star\|^2 = \|x_t - x_\star\|^2 - 2\frac{\gamma}{M(y_t)}\,\langle \nabla f(y_t), x_t - x_\star\rangle + \frac{\gamma^2}{M(y_t)^2}\,\|\nabla f(y_t)\|^2 .$$

Convexity at y_t gives $\langle \nabla f(y_t), y_t - x_\star\rangle \geq f(y_t) - f_\star$, and convexity at x_t gives $f(y_t) \geq f(x_t) + \langle \nabla f(x_t), y_t - x_t\rangle = f(x_t) + \rho\,\|\nabla f(x_t)\|^2 / M(x_t)$. Since $x_t - x_\star = (y_t - x_\star) - \rho\nabla f(x_t)/M(x_t)$,

$$\begin{aligned} \langle \nabla f(y_t), x_t - x_\star\rangle &\geq f(y_t) - f_\star - \frac{\rho}{M(x_t)}\,\langle \nabla f(y_t), \nabla f(x_t)\rangle \\ &\geq f(x_t) - f_\star - \frac{\rho}{M(x_t)}\,\langle \nabla f(y_t) - \nabla f(x_t), \nabla f(x_t)\rangle \\ &\geq f(x_t) - f_\star - \frac{\rho(e^{\rho}-1)}{M(x_t)}\,\|\nabla f(x_t)\|^2 \geq \Big(1 - \frac{\rho(e^{\rho}-1)}{2\log 2 - 1}\Big)\big(f(x_t) - f_\star\big), \end{aligned}$$

where we used the Cauchy–Schwarz inequality with Lemma B.4, and then Lemma B.3 in the form $\|\nabla f(x_t)\|^2 / M(x_t) \leq (f(x_t) - f_\star)/(2\log 2 - 1)$. Moreover, by Lemmas B.3 and B.4,

$$\frac{\|\nabla f(y_t)\|^2}{M(y_t)} \leq \frac{e^{2\rho}\,\|\nabla f(x_t)\|^2}{e^{-\rho}M(x_t)} = e^{3\rho}\frac{\|\nabla f(x_t)\|^2}{M(x_t)} \leq \frac{e^{3\rho}}{2\log 2 - 1}\big(f(x_t) - f_\star\big).$$

Substituting both bounds into the expansion,

$$\|x_{t+1} - x_\star\|^2 \leq \|x_t - x_\star\|^2 - 2\frac{\gamma}{M(y_t)}\Big[1 - \frac{\rho(e^{\rho}-1)}{2\log 2 - 1} - \frac{\gamma e^{3\rho}}{2(2\log 2 - 1)}\Big]\big(f(x_t) - f_\star\big),$$

and the bracket equals c_0. The last claim follows since $f(x_t) - f_\star \geq 0$. □

The next two lemmas concern the loss scale M_H of (4) only.

Lemma B.7. Under (4), for every $x \in \mathbb{R}^d$,

$$\|\nabla f(x)\|^2 \leq 2H_0\big(f(x) - f_\star\big) + H_1\big(f(x) - f_\star\big)^2, \qquad \sqrt{H_1}\,\|\nabla f(x)\| \leq M_H(x). \tag{29}$$

Proof. If $\nabla f(x) = 0$ there is nothing to prove. Suppose first $H_1 = 0$. If $H_0 = 0$, then $\nabla^2 f \equiv 0$, so f is affine and, being bounded below, constant, whence $\nabla f(x) = 0$. If $H_0 > 0$, then f is H_0-smooth and the descent lemma at the point $x - \nabla f(x)/H_0$ gives $f_\star \leq f(x - \nabla f(x)/H_0) \leq f(x) - \|\nabla f(x)\|^2/(2H_0)$, which is the first claim; the second is trivial.

Now let $H_1 > 0$. We compare the loss along the negative gradient direction with the solution of a scalar differential equation, using $f \geq f_\star$ to bound $\|\nabla f(x)\|$. Define, for $t \geq 0$, the function

$$\psi(t) := f\Big(x - t\,\frac{\nabla f(x)}{\|\nabla f(x)\|}\Big) - f_\star + \frac{H_0}{H_1}.$$

Then $\psi \geq H_0/H_1$, $\psi(0) = f(x) - f_\star + H_0/H_1$, $\psi'(0) = -\|\nabla f(x)\|$, and, by (4),

$$\psi''(t) \leq \left\|\nabla^2 f\Big(x - t\,\frac{\nabla f(x)}{\|\nabla f(x)\|}\Big)\right\| \leq H_0 + H_1\Big(f\Big(x - t\,\frac{\nabla f(x)}{\|\nabla f(x)\|}\Big) - f_\star\Big) = H_1\psi(t).$$

Let q solve $q'' = H_1 q$ with $q(0) = \psi(0)$ and $q'(0) = \psi'(0)$, that is

$$q(t) = \psi(0)\cosh\big(\sqrt{H_1}\,t\big) - \frac{\|\nabla f(x)\|}{\sqrt{H_1}}\sinh\big(\sqrt{H_1}\,t\big).$$

The difference $\psi - q$ vanishes together with its derivative at $t = 0$ and satisfies $(\psi-q)'' - H_1(\psi-q) = \psi'' - H_1\psi \leq 0$, so by the variation-of-constants formula

$$\psi(t) - q(t) = \int_0^t \frac{\sinh(\sqrt{H_1}(t-s))}{\sqrt{H_1}}\big(\psi''(s) - H_1\psi(s)\big)\,ds \leq 0.$$

Hence $q(t) \geq \psi(t) \geq H_0/H_1$ for all $t \geq 0$. If $\|\nabla f(x)\| > \sqrt{H_1}\,\psi(0)$, then $q(t) \to -\infty$ as $t \to \infty$, a contradiction; thus $\|\nabla f(x)\| \leq \sqrt{H_1}\,\psi(0)$, and then $\inf_{t\geq 0} q(t) = \sqrt{\psi(0)^2 - \|\nabla f(x)\|^2}\,/H_1$. This infimum is at least H_0/H_1, so

$$\begin{aligned}\|\nabla f(x)\|^2 &\leq H_1\Big(\psi(0)^2 - \frac{H_0^2}{H_1^2}\Big) = H_1\Big(f(x) - f_\star + \frac{H_0}{H_1}\Big)^2 - \frac{H_0^2}{H_1}\\ &= H_1\big(f(x) - f_\star\big)^2 + 2H_0\big(f(x) - f_\star\big),\end{aligned}$$

which is the first claim. For the second, multiply the first claim by H_1:

$$H_1\,\|\nabla f(x)\|^2 \leq 2H_0H_1\big(f(x) - f_\star\big) + H_1^2\big(f(x) - f_\star\big)^2 \leq \big(H_0 + H_1(f(x) - f_\star)\big)^2. \square$$

Lemma B.8. Under (4), if $M_H(x_0) = 0$ for some $x_0 \in \mathbb{R}^d$, then $M_H \equiv 0$ and f is constant.

Proof. $M_H(x_0) = 0$ forces $H_0 = 0$ and $f(x_0) = f_\star$. If $H_1 = 0$, then $\nabla^2 f \equiv 0$, so f is affine and bounded below, hence constant, and $M_H \equiv 0$. Let $H_1 > 0$ and fix $y \in \mathbb{R}^d$. The function $s \mapsto M_H(x_0 + s(y - x_0)) = H_1\big(f(x_0 + s(y - x_0)) - f_\star\big)$ is nonnegative and vanishes at $s = 0$, and by Lemma B.7 its derivative satisfies

$$\begin{aligned}\frac{d}{ds}M_H\big(x_0 + s(y - x_0)\big) &= H_1\,\langle \nabla f(x_0 + s(y - x_0)), y - x_0\rangle\\ &\leq \sqrt{H_1}\,M_H\big(x_0 + s(y - x_0)\big)\,\|y - x_0\|.\end{aligned}$$

Lemma B.1 with $c = \sqrt{H_1}\,\|y - x_0\|$ gives $M_H(x_0 + s(y - x_0)) \leq M_H(x_0)e^{\sqrt{H_1}s\|y-x_0\|} = 0$ on $[0, 1]$, so $M_H(y) = 0$, that is $f(y) = f_\star$. As y was arbitrary, f is constant. $\square$

We close this section with the elementary sequence lemma that turns a quadratic decrease of the gap into a $1/T$ rate.

Lemma B.9. Let $(x_t)_{t\geq 0}$ be a sequence such that $f(x_t) - f_\star$ is nonincreasing in t and

$$f(x_{t+1}) - f_\star \leq \big(f(x_t) - f_\star\big) - \beta\big(f(x_t) - f_\star\big)^2 \qquad \text{for all } t \geq 0,$$

where $\beta > 0$. Then $f(x_T) - f_\star \leq 1/(\beta T)$ for every $T \geq 1$.

Proof. If $f(x_T) = f_\star$ there is nothing to prove. Otherwise $f(x_t) - f_\star \geq f(x_T) - f_\star > 0$ for all $t \leq T$, and for $t < T$,

$$\frac{1}{f(x_{t+1}) - f_\star} - \frac{1}{f(x_t) - f_\star} = \frac{\big(f(x_t) - f_\star\big) - \big(f(x_{t+1}) - f_\star\big)}{\big(f(x_t) - f_\star\big)\big(f(x_{t+1}) - f_\star\big)} \geq \beta\,\frac{f(x_t) - f_\star}{f(x_{t+1}) - f_\star} \geq \beta.$$

Summing over $t < T$ gives $1/(f(x_T) - f_\star) \geq 1/(f(x_0) - f_\star) + \beta T \geq \beta T$. $\square$

C Proofs for Section 2

In this section, we present the proofs of the results of Section 2. We first verify the distortion bounds (6) and (7) for SAM and USAM. We then prove Proposition 2.1 by direct computation on the two one-dimensional examples, Theorem 2.2 by exhibiting two one-dimensional objectives in the class (3), and Proposition 2.4 using the lemmas of Appendix B. For completeness, we restate each result before its proof.

Verification of (6) and (7). The displacement is $d = r\nabla f(x)/\|\nabla f(x)\|$ for SAM and $d = \rho\nabla f(x)$ for USAM, so $\|d\| = r$ and $\|d\| = \rho\|\nabla f(x)\|$ respectively. Under (2), $\|\nabla f(x+d) - \nabla f(x)\| \le L\|d\|$, and dividing by $\|\nabla f(x)\|$ gives (6). Under (3) with $L_1 > 0$, the scale M_L is L_1-admissible by Proposition 2.4(ii), which is proved below, so the first inequality in Lemma B.2 gives

$$\|\nabla f(x+d) - \nabla f(x)\| \le \big(L_0 + L_1\|\nabla f(x)\|\big)\,\frac{e^{L_1\|d\|} - 1}{L_1}.$$

Inserting the two values of $\|d\|$ and dividing by $\|\nabla f(x)\|$ gives the two bounds in (7).

Proposition 2.1 (restated)**.**

(i) Let $f(x) = x^2/2$, which satisfies (2) with $L = 1$. Run SAM with $r > 0$ and $\eta \in (0, 2)$. For every $x \neq 0$,

$$D_{\mathrm{SAM}}(x) := D(x, r\,\mathrm{sign}(x)) = \frac{r}{|x|}, \qquad x_{t+1} = \big[1 - \eta\,(1 + D_{\mathrm{SAM}}(x_t))\big]\,x_t.$$

Let $a = \eta r/(2-\eta)$. Then $|x_{t+1}| < |x_t|$ if $|x_t| > a$, $|x_{t+1}| > |x_t|$ if $0 < |x_t| < a$, and $x_{t+1} = -x_t$ if $|x_t| = a$. Consequently $\pm a$ is a two-cycle, and no trajectory converges to the minimizer (unless it lands on it exactly).

(ii) Let $f(x) = x^4/4$, which satisfies (3) with $(L_0, L_1) = (4, 1)$. Run USAM with $\rho > 0$ and $\eta > 0$. For every $x \neq 0$,

$$D_{\mathrm{USAM}}(x) := D(x, \rho x^3) = (1+\rho x^2)^3 - 1, \qquad x_{t+1} = \big[1 - \eta\,(1 + D_{\mathrm{USAM}}(x_t))\,x_t^2\big]\,x_t.$$

Consequently, if $|x_0| > (2/(\eta\rho^3))^{1/8}$, then x_{t+1} has the opposite sign of x_t and $|x_{t+1}| > |x_t|$ for every t, and $|x_t| \to \infty$.

Proof of Proposition 2.1. *(i)* Here $\nabla f(x) = x$, so for $x \neq 0$ the perturbed point of SAM is $y = x + r\,\mathrm{sign}(x)$, and since $r\,\mathrm{sign}(x) = (r/|x|)\,x$,

$$D_{\mathrm{SAM}}(x) = \frac{|y - x|}{|x|} = \frac{r}{|x|}, \qquad x^+ = x - \eta y = x - \eta\Big(1 + \frac{r}{|x|}\Big)x = \big[1 - \eta\,(1 + D_{\mathrm{SAM}}(x))\big]\,x.$$

Write $c(x) := 1 - \eta - \eta r/|x|$ for the factor in brackets, so that $|x^+| = |c(x)|\,|x|$. Since $\eta > 0$ we have $c(x) < 1$, and since $\eta < 2$,

$$c(x) > -1 \iff 2 - \eta > \frac{\eta r}{|x|} \iff |x| > \frac{\eta r}{2-\eta} = a.$$

Hence $|c(x)| < 1$ if $|x| > a$, $|c(x)| > 1$ if $0 < |x| < a$, and $c(x) = -1$ if $|x| = a$, which are the three claims. In particular $a \mapsto -a \mapsto a$, so $\pm a$ is a two-cycle. Finally, suppose $x_t \neq 0$ for all t and $x_t \to 0$. Then $0 < |x_t| < a$ for all large t, so $|x_t|$ is eventually strictly increasing, which contradicts $x_t \to 0$.

(ii) Here $f'(x) = x^3$ and $f''(x) = 3x^2$. For $u \ge 0$ the function $3u^2 - u^3$ attains its maximum 4 at $u = 2$, so $3x^2 \le 4 + |x|^3 = 4 + |f'(x)|$ and (3) holds with $(L_0, L_1) = (4, 1)$. The perturbed point of USAM is $y = x + \rho x^3 = x(1 + \rho x^2)$, so that $y^3 = (1+\rho x^2)^3 x^3$ and, for $x \neq 0$,

$$D_{\mathrm{USAM}}(x) = \frac{|y^3 - x^3|}{|x|^3} = (1 + \rho x^2)^3 - 1, \qquad x^+ = x - \eta y^3 = \big[1 - \eta\,(1 + D_{\mathrm{USAM}}(x))\,x^2\big]\,x.$$

If $|x| > (2/(\eta\rho^3))^{1/8}$, then $\eta(1+\rho x^2)^3 x^2 \ge \eta\rho^3 x^8 > 2$, so the factor in brackets is less than -1: x^+ has the opposite sign of x and $|x^+| > |x|$. In particular $|x^+| > (2/(\eta\rho^3))^{1/8}$ again, so by induction the sign alternates and $|x_t|$ increases strictly for every t. Moreover, $|x_{t+1}| \ge (\eta\rho^3 x_t^8 - 1)|x_t| \ge (\eta\rho^3 x_0^8 - 1)|x_t|$ with $\eta\rho^3 x_0^8 - 1 > 1$, so $|x_t| \to \infty$. □

Theorem 2.2 (restated). Let $L_0 > 0$, $L_1 > 0$, and consider the perturbation given by a nonnegative collinear rule $d(x) = \alpha(\|\nabla f(x)\|)\nabla f(x)$. Suppose that for every one-dimensional objective satisfying (3) and every point with nonzero gradient, $D(x, d(x)) \le \delta$. Then, for every $s > 0$,

$$\alpha(s) \le \min\left\{\frac{\delta}{L_0}, \frac{\log(1+\delta)}{L_1 s}\right\}. \tag{10}$$

Proof of Theorem 2.2. Fix $s > 0$. Consider first $f_q(x) = L_0 x^2/2$. It is bounded below with $f_q'' \equiv L_0 \le L_0 + L_1|f_q'(x)|$, so it satisfies (3). At the point $x = s/L_0$ we have $f_q'(x) = s$, the perturbation is $d(x) = \alpha(s)s$, and

$$D\big(x, d(x)\big) = \frac{|f_q'(x + d(x)) - f_q'(x)|}{|f_q'(x)|} = \frac{L_0\alpha(s)s}{s} = L_0\alpha(s).$$

The hypothesis $D(x, d(x)) \le \delta$ gives $\alpha(s) \le \delta/L_0$.

Next consider $f_e(x) = e^{L_1 x}/L_1^2$. It is positive, hence bounded below, with $f_e'(x) = e^{L_1 x}/L_1 > 0$ and $f_e''(x) = e^{L_1 x} = L_1 f_e'(x) \le L_0 + L_1|f_e'(x)|$, so it satisfies (3). At the point $x = \log(L_1 s)/L_1$ we have $f_e'(x) = s$, the perturbation is $d(x) = \alpha(s)s \ge 0$, and since $f_e'(x + d) = e^{L_1 d} f_e'(x)$,

$$D\big(x, d(x)\big) = \frac{f_e'(x + d(x)) - f_e'(x)}{f_e'(x)} = e^{L_1\alpha(s)s} - 1.$$

The hypothesis gives $e^{L_1\alpha(s)s} \le 1 + \delta$, that is $\alpha(s) \le \log(1+\delta)/(L_1 s)$. Together with the first bound this is (10). □

Proposition 2.4 (restated).

(i) Under (2), the constant scale $M \equiv L$ is admissible with $\kappa = 0$.

(ii) Under (3), the scale M_L is admissible with $\kappa = L_1$.

(iii) Under (4), either $M_H \equiv 0$ and f is constant, or $M_H > 0$ everywhere and the scale M_H is admissible with $\kappa = \sqrt{H_1}$.

(iv) Under (3), for every $x \in \mathbb{R}^d$,

$$L_0 + L_1\|\nabla f(x)\| \le \frac{3}{2}L_0 + \frac{3L_1^2}{2\,(2\log 2 - 1)}\big(f(x) - f_\star\big). \tag{12}$$

In particular, (3) implies (4) with the constants $H_0 = 3L_0/2$ and $H_1 = 3L_1^2/(2(2\log 2 - 1))$.

Proof of Proposition 2.4. *(i)* With $M \equiv L$ and $\kappa = 0$, the three conditions in (11) read $\|\nabla^2 f(x)\| \le L$, $L \le L$, and $0 \le L$.

(ii) The first condition is (3), and the third holds with $\kappa = L_1$ because $L_1\|\nabla f(x)\| \le L_0 + L_1\|\nabla f(x)\|$. For the second, fix x, y. Since ∇f is continuously differentiable, the function $s \mapsto M_L(x+s(y-x)) = L_0 + L_1\|\nabla f(x + s(y-x))\|$ is Lipschitz on $[0,1]$, hence absolutely continuous, and at every point of differentiability

$$\begin{aligned}\frac{d}{ds}M_L\big(x + s(y-x)\big) &\le L_1\big\|\nabla^2 f(x + s(y-x))\,(y-x)\big\| \le L_1\big\|\nabla^2 f(x + s(y-x))\big\|\,\|y - x\| \\ &\le L_1 M_L\big(x + s(y-x)\big)\,\|y - x\|.\end{aligned}$$

Lemma B.1 with $c = L_1\|y - x\|$ gives $M_L(y) \le M_L(x)e^{L_1\|y-x\|}$.

(iii) If M_H vanishes somewhere, Lemma B.8 gives $M_H \equiv 0$ and f constant. Otherwise $M_H > 0$ everywhere. The first condition is (4), and the third, with $\kappa = \sqrt{H_1}$, is Lemma B.7. For the second, fix x, y; by Lemma B.7,

$$\frac{d}{ds}M_H\big(x + s(y-x)\big) = H_1\,\langle\nabla f(x + s(y-x)), y - x\rangle \le H_1\,\|\nabla f(x + s(y-x))\|\,\|y - x\|$$

$$\leq \sqrt{H_1}\, M_H\big(x + s(y-x)\big)\, \|y - x\|\,,$$

and Lemma B.1 with $c = \sqrt{H_1}\,\|y-x\|$ gives $M_H(y) \leq M_H(x) e^{\sqrt{H_1}\|y-x\|}$.

(iv) By (ii), Lemma B.3 applies to M_L and gives $(2\log 2 - 1)\,\|\nabla f(x)\|^2 \leq (f(x) - f_\star)\big(L_0 + L_1\,\|\nabla f(x)\|\big)$, that is

$$\|\nabla f(x)\|^2 - \frac{L_1(f(x)-f_\star)}{2\log 2 - 1}\,\|\nabla f(x)\| - \frac{L_0(f(x)-f_\star)}{2\log 2 - 1} \leq 0.$$

Since $s^2 \leq A + Bs$ with $A, B, s \geq 0$ implies $s \leq \big(B + \sqrt{B^2 + 4A}\big)/2 \leq B + \sqrt{A}$,

$$\|\nabla f(x)\| \leq \frac{L_1(f(x)-f_\star)}{2\log 2 - 1} + \sqrt{\frac{L_0(f(x)-f_\star)}{2\log 2 - 1}}.$$

Therefore, using $\sqrt{AB} \leq (A+B)/2$ with $A = L_0$ and $B = L_1^2(f(x)-f_\star)/(2\log 2 - 1)$,

$$\begin{aligned} L_0 + L_1\,\|\nabla f(x)\| &\leq L_0 + \frac{L_1^2(f(x)-f_\star)}{2\log 2 - 1} + \sqrt{L_0 \cdot \frac{L_1^2(f(x)-f_\star)}{2\log 2 - 1}} \\ &\leq \frac{3}{2}L_0 + \frac{3L_1^2}{2(2\log 2 - 1)}\big(f(x) - f_\star\big), \end{aligned}$$

which is (12). Combined with (3), this bounds $\big\|\nabla^2 f(x)\big\|$ by the same right-hand side, which is (4) with the stated constants. □

Theorem 2.5 (restated)**.** Let M be κ-admissible, let $\rho \geq 0$, $\gamma > 0$, and run CS-SAM. Then every iteration with $\nabla f(x_t) \neq 0$ satisfies

$$e^{-\rho} M(x_t) \leq M(y_t) \leq e^{\rho} M(x_t), \qquad \frac{\|\nabla f(y_t) - \nabla f(x_t)\|}{\|\nabla f(x_t)\|} \leq e^{\rho} - 1.$$

In particular, $D_{\text{CS-SAM}} \leq e^{\rho} - 1$.

Proof of Theorem 2.5. The two claims are the inequalities in (25) of Lemma B.4, which holds for every κ-admissible scale M and every $\rho \geq 0$; since $\nabla f(x_t) \neq 0$, dividing the second inequality by $\|\nabla f(x_t)\|$ gives the distortion bound. □

D PROOFS FOR SECTION 3

In this section, we present the proofs of Theorem 3.1 in Appendix D.1 and of Theorem 3.2 in Appendix D.2; for completeness, each theorem is restated before its proofs.

D.1 PROOFS FOR SECTION 3.1: THE (L_0, L_1) CURVATURE SCALE

Throughout this subsection f satisfies (3) and $M = M_L$, which is admissible with $\kappa = L_1$ by Proposition 2.4(ii), so Lemmas B.2 to B.6 apply with $M(x_t) = L_0 + L_1 \|\nabla f(x_t)\|$ and $M(y_t) = L_0 + L_1 \|\nabla f(y_t)\|$. We use repeatedly that the map $s \mapsto s^2/(L_0 + L_1 s)$ is increasing on $[0, \infty)$, since its derivative is $(2L_0 s + L_1 s^2)/(L_0 + L_1 s)^2 \geq 0$. In this setting Lemma B.5 reads

$$f(x_{t+1}) - f_\star \leq \big(f(x_t) - f_\star\big) - \gamma e^{-\rho} \chi(\rho, \gamma) \, \frac{\|\nabla f(x_t)\|^2}{L_0 + L_1 \|\nabla f(x_t)\|}. \tag{30}$$

Theorem 3.1 (restated)**.** Let f satisfy (3), let $\rho \geq 0$ and $\gamma > 0$ satisfy $\chi(\rho, \gamma) > 0$, and let $(x_t)_{t \geq 0}$ be generated by CS-SAM with $M = M_L$. Then, for every $T \geq 1$, the following hold:

(i) *(Nonconvex rate)*

$$\min_{0 \leq t < T} \|\nabla f(x_t)\|^2 \leq \left(\sqrt{\frac{e^\rho L_0 \big(f(x_0) - f_\star\big)}{\gamma \chi(\rho, \gamma) \, T}} + \frac{e^\rho L_1 \big(f(x_0) - f_\star\big)}{\gamma \chi(\rho, \gamma) \, T} \right)^2. \tag{16}$$

(ii) *(Convex rate)* If in addition f is convex with a minimizer $x_\star$ and $c_0 \geq 0$, then

$$f(x_T) - f_\star \leq \frac{e^\rho \Big(L_0 \|x_0 - x_\star\|^2 + L_1 \|x_0 - x_\star\| \big(f(x_0) - f_\star\big) \Big)}{\gamma \chi(\rho, \gamma) \, T}. \tag{17}$$

(iii) *(PL rate)* If in addition f satisfies the μ-PL condition (15), then

$$f(x_T) - f_\star \leq \left(1 - \frac{2\mu \gamma e^{-\rho} \chi(\rho, \gamma)}{L_0 + L_1 \sqrt{2\mu \big(f(x_0) - f_\star\big)}} \right)^T \big(f(x_0) - f_\star\big). \tag{18}$$

Proof of Theorem 3.1. (i) Summing (30) over $t < T$ gives

$$\gamma e^{-\rho} \chi(\rho, \gamma) \sum_{t<T} \frac{\|\nabla f(x_t)\|^2}{L_0 + L_1 \|\nabla f(x_t)\|} \leq f(x_0) - f(x_T) \leq f(x_0) - f_\star.$$

Since the map $s \mapsto s^2/(L_0 + L_1 s)$ is increasing, each summand is bounded below by the value of this map at the smallest of the gradient norms, hence

$$\Big(\min_{t<T} \|\nabla f(x_t)\| \Big)^2 \leq \frac{e^\rho \big(f(x_0) - f_\star\big)}{\gamma \chi(\rho, \gamma) \, T} \Big(L_0 + L_1 \min_{t<T} \|\nabla f(x_t)\| \Big).$$

Since $s^2 \leq A + Bs$ with $A, B, s \geq 0$ implies $s \leq B + \sqrt{A}$, we get

$$\min_{t<T} \|\nabla f(x_t)\| \leq \sqrt{\frac{e^\rho L_0 \big(f(x_0) - f_\star\big)}{\gamma \chi(\rho, \gamma) \, T}} + \frac{e^\rho L_1 \big(f(x_0) - f_\star\big)}{\gamma \chi(\rho, \gamma) \, T},$$

and squaring gives (16).

(ii) Since $c_0 \geq 0$, Lemma B.6 gives $\|x_t - x_\star\| \leq \|x_0 - x_\star\|$ for all t. Convexity and the Cauchy–Schwarz inequality give

$$f(x_t) - f_\star \leq \langle \nabla f(x_t), x_t - x_\star \rangle \leq \|\nabla f(x_t)\| \, \|x_0 - x_\star\|.$$

Since the map $s \mapsto s^2/(L_0 + L_1 s)$ is increasing and $f(x_t) - f_\star \le f(x_0) - f_\star$,

$$\frac{\|\nabla f(x_t)\|^2}{L_0 + L_1 \|\nabla f(x_t)\|} \ge \frac{\big(f(x_t) - f_\star\big)^2}{L_0 \|x_0 - x_\star\|^2 + L_1 \|x_0 - x_\star\| \big(f(x_t) - f_\star\big)} \ge \frac{\big(f(x_t) - f_\star\big)^2}{L_0 \|x_0 - x_\star\|^2 + L_1 \|x_0 - x_\star\| \big(f(x_0) - f_\star\big)}.$$

Inserting this into (30) gives

$$f(x_{t+1}) - f_\star \le \big(f(x_t) - f_\star\big) - \frac{\gamma e^{-\rho} \chi(\rho,\gamma)}{L_0 \|x_0 - x_\star\|^2 + L_1 \|x_0 - x_\star\| \big(f(x_0) - f_\star\big)} \big(f(x_t) - f_\star\big)^2.$$

Since $f(x_t) - f_\star$ is nonincreasing by Lemma B.5, Lemma B.9 applies and gives (17).

(iii) If $f(x_0) = f_\star$ all iterates are minimizers and there is nothing to prove, so let $f(x_0) > f_\star$. The PL condition (15) gives $\|\nabla f(x_t)\| \ge \sqrt{2\mu(f(x_t) - f_\star)}$, so by the monotonicity of $s \mapsto s^2/(L_0 + L_1 s)$, (30), and $f(x_t) - f_\star \le f(x_0) - f_\star$,

$$f(x_{t+1}) - f_\star \le \big(f(x_t) - f_\star\big) - \gamma e^{-\rho} \chi(\rho,\gamma) \frac{2\mu\big(f(x_t) - f_\star\big)}{L_0 + L_1 \sqrt{2\mu(f(x_t) - f_\star)}} \le \Big(1 - \frac{2\mu\gamma e^{-\rho} \chi(\rho,\gamma)}{L_0 + L_1 \sqrt{2\mu(f(x_0) - f_\star)}}\Big) \big(f(x_t) - f_\star\big).$$

Applying this with $t = 0$ and using $f(x_1) - f_\star \ge 0$ and $f(x_0) - f_\star > 0$ shows that the factor in parentheses is nonnegative, and it is less than one since $\mu\gamma e^{-\rho}\chi(\rho,\gamma) > 0$. Induction on t then gives (18).

□

D.2 Proofs for Section 3.2: the (H_0, H_1) curvature scale

Throughout this subsection f satisfies (4), is nonconstant, and $M = M_H$, which is positive and admissible with $\kappa = \sqrt{H_1}$ by Proposition 2.4(iii), so Lemmas B.2 to B.6 apply with $M(x_t) = H_0 + H_1(f(x_t) - f_\star)$ and $M(y_t) = H_0 + H_1(f(y_t) - f_\star)$. Since $f(x_t) - f_\star$ is nonincreasing by Lemma B.5, we have $M(x_t) \le H_0 + H_1(f(x_0) - f_\star)$ for all t, and Lemma B.5 gives

$$\begin{aligned} f(x_{t+1}) - f_\star &\le \big(f(x_t) - f_\star\big) - \gamma e^{-\rho} \chi(\rho,\gamma) \frac{\|\nabla f(x_t)\|^2}{H_0 + H_1(f(x_t) - f_\star)} \\ &\le \big(f(x_t) - f_\star\big) - \gamma e^{-\rho} \chi(\rho,\gamma) \frac{\|\nabla f(x_t)\|^2}{H_0 + H_1(f(x_0) - f_\star)}. \end{aligned} \tag{31}$$

Note that the second inequality in (31) has a constant coefficient in front of $\|\nabla f(x_t)\|^2$, which is why the H_1 contribution enters the rates of Theorem 3.2 at the leading order, in contrast to Theorem 3.1.

Theorem 3.2 (restated). Let f satisfy (4) and be nonconstant, let $\rho \ge 0$ and $\gamma > 0$ satisfy $\chi(\rho,\gamma) > 0$, and let $(x_t)_{t\ge 0}$ be generated by CS-SAM with $M = M_H$. Then, for every $T \ge 1$, the following hold:

(i) *(Nonconvex rate)*

$$\min_{0 \le t < T} \|\nabla f(x_t)\|^2 \le \frac{e^\rho \big(H_0 + H_1(f(x_0) - f_\star)\big)\big(f(x_0) - f_\star\big)}{\gamma \chi(\rho,\gamma)\, T}. \tag{19}$$

(ii) *(Convex rate)* If in addition f is convex with a minimizer $x_\star$ and $c_0 \ge 0$, then

$$f(x_T) - f_\star \le \frac{e^\rho \big(H_0 + H_1(f(x_0) - f_\star)\big) \|x_0 - x_\star\|^2}{\gamma \chi(\rho,\gamma)\, T}. \tag{20}$$

(iii) *(PL rate)* If in addition f satisfies the μ-PL condition (15), then

$$f(x_T) - f_\star \le \left(1 - \frac{2\mu\gamma e^{-\rho}\chi(\rho,\gamma)}{H_0 + H_1\big(f(x_0) - f_\star\big)}\right)^T \big(f(x_0) - f_\star\big). \tag{21}$$

Proof of Theorem 3.2. (i) Summing the second inequality in (31) over $t < T$ gives

$$\frac{\gamma e^{-\rho}\chi(\rho,\gamma)}{H_0 + H_1(f(x_0) - f_\star)} \sum_{t<T} \|\nabla f(x_t)\|^2 \le f(x_0) - f(x_T) \le f(x_0) - f_\star,$$

and the minimum over $t < T$ is at most the average, which gives (19).

(ii) Since $c_0 \ge 0$, Lemma B.6 gives $\|x_t - x_\star\| \le \|x_0 - x_\star\|$, and convexity gives $\|\nabla f(x_t)\| \ge (f(x_t) - f_\star)/\|x_0 - x_\star\|$ as in the proof of Theorem 3.1(ii). Inserting this into the second inequality in (31),

$$f(x_{t+1}) - f_\star \le \big(f(x_t) - f_\star\big) - \frac{\gamma e^{-\rho}\chi(\rho,\gamma)}{\|x_0 - x_\star\|^2\big(H_0 + H_1(f(x_0) - f_\star)\big)}\big(f(x_t) - f_\star\big)^2,$$

and Lemma B.9 gives (20).

(iii) Let $f(x_0) > f_\star$, the other case being trivial. The PL condition (15) and the second inequality in (31) give

$$f(x_{t+1}) - f_\star \le \Big(1 - \frac{2\mu\gamma e^{-\rho}\chi(\rho,\gamma)}{H_0 + H_1(f(x_0) - f_\star)}\Big)\big(f(x_t) - f_\star\big).$$

As in the proof of Theorem 3.1(iii), the case $t = 0$ shows that the factor in parentheses lies in $[0, 1)$, and induction gives (21).

□

E EXPERIMENTAL DETAILS

In this section, we give the details needed to reproduce the experiments of Section 4. Appendix E.1 specifies the one-dimensional verification experiments of Section 4.1: the update rules, the curvature scales and the constants that make them valid everywhere, the stopping rules, and the evaluation of the distortion. Appendix E.2 gives, for the deep-network experiments of Section 4.2, the training protocol, the minibatch curvature scales and their fixed coefficients, the tuning campaign with the complete grid of every method, the selection rule, and the selected configuration of every method in every model–dataset setting (Table 4).

The pseudocode for CS-SAM is given in Algorithm 1.

Algorithm 1 Curvature-Scaled SAM (CS-SAM).

Require: starting point $x_0 \in \mathbb{R}^d$; perturbation multiplier $\rho \geq 0$; step multiplier $\gamma > 0$; admissible curvature scale M (Definition 2.3)

for $t = 0, 1, \ldots$ **do**
  if $\nabla f(x_t) = 0$ **then**
    return x_t ▷ *stationary point reached; the method stops*
  end if
  $y_t \leftarrow x_t + \dfrac{\rho}{M(x_t)} \nabla f(x_t)$
  $x_{t+1} \leftarrow x_t - \dfrac{\gamma}{M(y_t)} \nabla f(y_t)$
end for

E.1 VERIFICATION EXPERIMENTS OF SECTION 4.1

The experiments of Section 4.1 are deterministic and one-dimensional, with analytic function values and gradients, so there are no datasets, minibatches, random seeds, or hyperparameter searches. All computations use double precision.

Update rules and scales. The three methods are run exactly as stated in (SAM), (USAM), and (CS-SAM), with $\nabla f = f'$ and the norm equal to the absolute value. Both objectives attain $f_\star = 0$ at $x = 0$, and the two scales are

$$M_L(x) = L_0 + L_1|f'(x)|, \qquad M_H(x) = H_0 + H_1\big(f(x) - f_\star\big).$$

For the quadratic $f(x) = x^2/2$ we have $f'' \equiv 1$ and take $(L_0, L_1) = (H_0, H_1) = (1, 0)$, so that $M_L = M_H \equiv 1$ and a single CS-SAM curve is shown. For the quartic $f(x) = x^4/4$ we have $f''(x) = 3x^2$ and take $(L_0, L_1) = (4, 1)$ and $(H_0, H_1) = (3, 3)$, that is, $M_L(x) = 4 + |x|^3$ and $M_H(x) = 3 + 3x^4/4$.

Parameter regime. The common choice $\rho = \gamma = 1/8$ for CS-SAM lies in the regime of Theorems 3.1 and 3.2. With the constants of (14), $2 - e^{1/8} \approx 0.867$ and $\gamma\omega(\gamma)e^{3\rho} = \frac{1}{8}\,\omega(\frac{1}{8})\,e^{3/8} \approx 0.095$, so that $\chi(1/8, 1/8) \approx 0.772 > 0$; moreover $\rho(e^\rho - 1)/(2\log 2 - 1) \approx 0.043$ and $\gamma e^{3\rho}/(2(2\log 2 - 1)) \approx 0.235$, so that $c_0 \approx 0.722 \geq 0$. Since both objectives are convex with the minimizer $x_\star = 0$, the convergence of CS-SAM in Figures 1a and 1b is the one guaranteed by Theorem 3.1(ii) with $M = M_L$ and by Theorem 3.2(ii) with $M = M_H$, and the bound $e^\rho - 1$ in Figures 1c and 1d is that of Theorem 2.5, which holds for every $\rho \geq 0$.

Two-cycle and divergence (Figures 1a and 1b). A run stops when $|f'(x)| \leq 10^{-12}$, when an update no longer changes the double-precision iterate or when an iterate or a perturbed point has absolute value at least 10^{12}. On the quadratic, SAM uses $r = 1$ and $\eta = 0.2$, and CS-SAM uses $\rho = \gamma = 1/8$, both from $x_0 = 3$ with at most 500 updates. The SAM recurrence is $x_{t+1} = 0.8x_t - 0.2\,\mathrm{sign}(x_t)$, whose two-cycle has amplitude $a = \eta r/(2 - \eta) = 1/9$, marked by the dashed lines, and the CS-SAM recurrence is $x_{t+1} = (1 - \gamma(1 + \rho))x_t = \frac{55}{64}x_t$. SAM completes the 500 updates, whereas CS-SAM reaches the gradient tolerance after 190 updates. On the quartic, USAM uses $\rho = 1/8$ and $\eta = 1/32$ with at most 20 updates, and CS-SAM uses $\rho = \gamma = 1/8$ with either scale and at most 10,000 updates, all from $x_0 = 4$. For USAM, the first perturbed point is $y_0 = 4 + \frac{1}{8} \cdot 4^3 = 12$, so that $x_1 = 4 - \frac{1}{32} \cdot 12^3 = -50$, then $x_2 \approx 1.2 \times 10^{11}$, and the next perturbed point exceeds the cutoff; the

trace therefore ends after two completed updates, and the plotted curve consists of its three recorded points. Both CS-SAM traces complete the 10,000 updates. Figure 1b shows $|x_t|$ against the number $2t$ of gradient evaluations, two per completed update, over the first 50 updates. Note that the cutoff is a numerical termination rule; the divergence itself is the statement of Proposition 2.1(ii), whose threshold $(2/(\eta\rho^3))^{1/8} \approx 3.67$ is exceeded by $x_0 = 4$.

Distortion (Figures 1c and 1d). These figures evaluate independent points rather than trajectories. At each nonstationary x we form the perturbed point y of the method and record the gradient norm $s = |f'(x)|$ and the distortion $D = |f'(y) - f'(x)|/|f'(x)|$ of (5); no outer step enters. The perturbations are those of SAM with $r = 1/8$ on the quadratic, of USAM with $\rho = 1/8$ on the quartic, and of CS-SAM with $\rho = 1/8$ on both, with each scale separately on the quartic. Note that the SAM radius is 1 in the two-cycle experiment and $1/8$ here; the two are separate experiments, and a common numerical value of the parameters does not mean equal perturbation lengths across methods. The dashed line is the bound $e^{\rho} - 1 = e^{1/8} - 1 \approx 0.133$ of Theorem 2.5.

E.2 Deep-network experiments of Section 4.2

Training protocol. All runs train ResNet-20 or ResNet-32 with batch normalization on CIFAR-10 or CIFAR-100 for 200 epochs with batch size 128 and the cross-entropy loss, with random-crop and horizontal-flip augmentation and no cutout. Momentum and weight decay are zero for every method. The test accuracy is evaluated after every epoch. Every method computes two minibatch gradients per iteration, one at the iterate and one at the perturbed point, so that equal numbers of epochs correspond to equal numbers of gradient evaluations. The runs were performed on four RTX 6000 Ada GPUs.

Learning-rate schedules. The schedule of SAM and USAM acts on the outer step η_t only; the radius r and the multiplier ρ are constant throughout training. With the constant schedule, $\eta_t \equiv \eta$. With the cosine schedule, the outer step is held fixed within each epoch and annealed from η towards 0 over the $E = 200$ epochs: during epoch $e \in \{0, \dots, E-1\}$,

$$\eta_t = \frac{\eta}{2}\Big(1 + \cos\frac{\pi e}{E}\Big), \tag{32}$$

so that $\eta_t = \eta$ during the first epoch and $\eta_t \approx 6 \times 10^{-5}\,\eta$ during the last. This is cosine annealing without restarts, warm-up, or a positive floor (Loshchilov & Hutter, 2017), applied once per epoch through the scheduler interface of PyTorch (Paszke et al., 2019).

Minibatch curvature scales. As stated in Section 4.2, the M_L scale of a minibatch is $L_0 + L_1\,\|\nabla f_i(x)\|$ and the M_H scale is $H_0 + H_1[f_i(x) - \ell_i^\star]_+$ with $\ell_i^\star = 0$, where f_i and ∇f_i are the minibatch loss and gradient. Note that multiplying (ρ, γ, L_0, L_1), or (ρ, γ, H_0, H_1), by one positive constant leaves CS-SAM unchanged, since the perturbation step $\rho/M(x_t)$ and the descent step $\gamma/M(y_t)$ are ratios. Only three of the four coefficients of each scale are therefore free, and we fix $L_1 = 1$ for the M_L scale and $H_0 = 1$ for the M_H scale, so that the tuned coefficients are (ρ, γ, L_0) and (ρ, γ, H_1) respectively. The coefficients are constant over training; CS-SAM uses no learning-rate schedule, since the perturbation and the descent step adapt through the scale.

Tuning protocol. The protocol has three stages, each run on all four model–dataset settings.

(i) *Baselines.* SAM and USAM, each with a constant and with a cosine annealing schedule, are tuned on seed 3107 over a 4×4 grid of learning rate and radius (multiplier): $\eta \in \{0.03, 0.1, 0.3, 1\}$ for all four baselines, $r \in \{0.01, 0.05, 0.1, 0.2\}$ for SAM, and $\rho \in \{0.1, 0.3, 1, 3\}$ for USAM. For the cosine annealing schedule (32), η is the initial learning rate.

(ii) *CS-SAM.* The M_L scale is tuned on seed 3107 over $\rho \in \{0.2, 0.4, 0.8, 1.6\}$, $\gamma \in \{0.2, 0.4, 0.8\}$, and $L_0 \in \{0.02, 0.5, 2\}$ with $L_1 = 1$, and the M_H scale over $\rho \in \{0.3, 1, 3\}$, $\gamma \in \{0.3, 1, 3\}$, and $H_1 \in \{1, 3, 10, 30\}$ with $H_0 = 1$. This stage comprises 36 configurations per scale and setting, that is, 288 runs.

(iii) *Seeds.* For every method and setting, the configuration with the highest best test accuracy over the 200 epochs on seed 3107 is rerun with seeds 3407 and 4096. Table 3 reports the mean and the sample standard deviation over the three seeds of the best test accuracy over epochs.

Selected configurations. Table 4 lists, for every method and setting, the configuration selected in stages (i) and (ii) and rerun in stage (iii); these are the configurations behind the entries of Table 3.

Table 4: Configurations selected on seed 3107 by the best test accuracy over 200 epochs, for every method and setting of Table 3. For SAM and USAM, η is the constant learning rate or the initial learning rate of the cosine annealing schedule (32), r is the radius of SAM, and ρ is the multiplier of USAM. For CS-SAM, ρ and γ are the perturbation and step multipliers, L_0 is the constant term of the M_L scale $L_0 + L_1 \|\nabla f_i(x)\|$ with $L_1 = 1$, and H_1 is the loss coefficient of the M_H scale $H_0 + H_1[f_i(x) - \ell_i^\star]_+$ with $H_0 = 1$ and $\ell_i^\star = 0$.

Method	Parameters	ResNet-20 / CIFAR-10	ResNet-32 / CIFAR-10	ResNet-20 / CIFAR-100	ResNet-32 / CIFAR-100
SAM, constant lr	(η, r)	(0.3, 0.2)	(0.3, 0.2)	(0.1, 0.2)	(0.1, 0.2)
SAM, cosine	(η, r)	(1, 0.1)	(0.3, 0.2)	(0.3, 0.2)	(0.3, 0.2)
USAM, constant lr	(η, ρ)	(0.3, 0.3)	(0.3, 1)	(0.3, 1)	(0.3, 0.3)
USAM, cosine	(η, ρ)	(0.3, 0.3)	(0.3, 0.3)	(0.3, 0.3)	(0.3, 0.3)
CS-SAM with M_L	(ρ, γ, L_0)	(0.8, 0.4, 0.02)	(0.8, 0.4, 0.02)	(0.4, 0.2, 0.02)	(0.4, 0.4, 0.02)
CS-SAM with M_H	(ρ, γ, H_1)	(3, 3, 10)	(3, 1, 1)	(3, 3, 3)	(3, 3, 3)