PRECONDITIONED SHARPNESS-AWARE MINIMIZATION CAN EXPLOIT LARGE PERTURBATION RADII

Anonymous authors
Paper under double-blind review

ABSTRACT

When sharpness-aware minimization (SAM) is combined with Adam-type optimizers, its perturbation and update operate in different geometries. Preconditioned SAM (PSAM) (Singh et al., 2025) addresses this mismatch. We analyze PSAM in its own right and examine its practical use. We show that each PSAM step is standard SAM in scaled geometry when the perturbation and update share a preconditioner. This yields a stationarity bound and a fixed-metric PAC-Bayes bound. Our descent analysis under quasiconvexity introduces an admissible radius and suggests how PSAM's perturbation direction can allow larger radii. In practice, early preconditioner instability motivates a warm start for PSAM. Across vision and language tasks, warm-started PSAM performs best among the evaluated methods. In language modeling, the best-performing radius for warm-started PSAM is up to $25\times$ that for warm-started SAM, while other methods deteriorate at these radii. We observe larger radius gaps when the SAM and PSAM perturbation directions are less aligned, consistent with our geometric analysis.

1 INTRODUCTION

Flat minimizers are often linked to good generalization (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Neyshabur et al., 2017). To find them, sharpness-aware minimization (SAM) (Foret et al., 2021) uses the worst-case loss over a Euclidean ball, with a first-order approximation:

$$\max_{\|\epsilon\|_2\le\rho} f(x+\epsilon) = f(x) + \max_{\|\epsilon\|_2\le\rho} \big[f(x+\epsilon) - f(x)\big], \qquad \epsilon = \rho\frac{\nabla f(x)}{\|\nabla f(x)\|_2}. \tag{1}$$

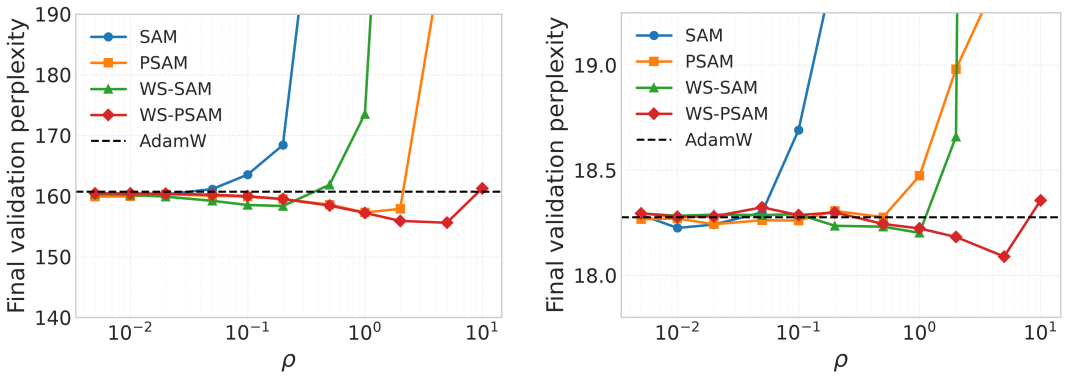
The two terms on the right are loss and sharpness, so minimizing their sum favors flat minimizers. Here, the perturbation radius ρ determines the size of the neighborhood over which sharpness is measured, thereby controlling how strongly SAM penalizes sharpness. Across a range of deep learning tasks in computer vision and natural language processing (NLP), SAM has been shown to improve generalization (Foret et al., 2021; Chen et al., 2022; Bahri et al., 2022).

When SAM is combined with Adam-type optimizers, commonly used in NLP, its Euclidean perturbation and preconditioned update operate in different geometries. To address this mismatch, Singh et al. (2025) propose preconditioned SAM (PSAM), which applies the optimizer's preconditioner to the perturbation. They show that PSAM can improve upon AdamW in language-modeling settings where standard SAM degrades performance. However, PSAM is studied alongside Functional SAM, a distinct method with which it can be combined, rather than as the sole focus of the work.

In this work, we take PSAM as our object of study. On language-modeling tasks, warm-started PSAM outperforms all compared methods, attaining its best performance at radii up to $25\times$ larger than the optimal radius of warm-started SAM (Figure 1). These radii are directly comparable because both SAM and PSAM use perturbations of Euclidean norm ρ. At these large radii, the other methods suffer substantial performance degradation or training failure. How can PSAM exploit such large radii to achieve superior performance?

We use scaled geometry to extend SAM's theoretical guarantees to PSAM. Through the notion of an admissible radius, we explain how preconditioning the perturbation direction can allow larger radii. In practice, we find that a warm start alleviates early-training instability and improves performance at large radii. Finally, we measure the alignment between the two perturbation directions, supporting our geometric explanation of PSAM's large-radius behavior.

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.



(a) GPT-2 (30M), WikiText-2.

(b) GPT-2 (124M), WikiText-103.

Figure 1: Warm-started PSAM achieves the best performance at large radii where the other methods deteriorate. Final validation perplexity versus ρ for (a) GPT-2 (30M) on WikiText-2 and (b) GPT-2 (124M) on WikiText-103 (lower is better), with the AdamW baseline. WS- denotes warm-started variants. Results are averaged over five seeds for GPT-2 (30M) and three seeds for GPT-2 (124M).

Our contributions are summarized as follows:

- We show that each PSAM step is standard SAM in scaled geometry when the perturbation and update share a preconditioner. We derive a stationarity bound and a fixed-metric PAC-Bayes bound for PSAM.
- We establish a descent property under quasiconvexity for SAM and PSAM, show that a geometry mismatch can invalidate it, and introduce an admissible radius to give a geometric explanation for why PSAM can admit larger radii.
- We identify early preconditioner drift that motivates warm-starting PSAM. Our experiments across vision and NLP tasks show that warm-started PSAM outperforms SAM, PSAM, and warm-started SAM. The relationship between perturbation alignment and the gap in best-performing radii is consistent with our geometric analysis.

2 Related Work

Analyses of SAM. SAM (Foret et al., 2021) has been studied from many angles (Kaddour et al., 2022; Wen et al., 2023; Möllenhoff & Khan, 2023; Bartlett et al., 2023; Dai et al., 2023; Moon et al., 2026; Kim et al., 2026). Its convergence has been analyzed (Si & Yun, 2023; Khanh et al., 2024; Oikonomou & Loizou, 2025), also with adaptive learning rates (Sun et al., 2024). Its generalization has also been studied (Foret et al., 2021; Chen et al., 2023); we extend the stationarity and PAC-Bayes bounds to PSAM (Section 5). The sharpness measured by SAM depends on the parameterization (Dinh et al., 2017), motivating alternatives to its Euclidean perturbation geometry.

Preconditioned perturbation. Many SAM variants use a preconditioned perturbation whose direction differs from the gradient. Zhang et al. (2025) group them into constraint preconditioning (CP), which holds its scaled norm fixed (Kwon et al., 2021; Kim et al., 2022; Yun & Yang, 2023), and objective preconditioning (OP), which holds its Euclidean norm fixed (Li & Giannakis, 2023; Jiang et al., 2023; Luo et al., 2025; Becker et al., 2025). Analyzing both under an SGD update, they note that CP can slow convergence and prefer OP for its stability. Singh et al. (2025) propose PSAM, an OP method that matches the perturbation and update geometries for Adam-type optimizers.

Warm start and large radius. Andriushchenko & Flammarion (2022) experiment with when to switch SAM on, and Zhou et al. (2025) find that switching late already recovers most of the generalization benefit. Park et al. (2026) switch early instead, and show that such a warm start prevents a failure mode of SAM at large radii: convergence to hallucinated minimizers. Warm starts and large radii also arise in Singh et al. (2025), who introduced PSAM: they apply a warm-up to keep PSAM from diverging in some runs, and find that preconditioned Functional SAM works well at large radii. What a warm start does for PSAM itself at large radii, however, is left open.

3 PRELIMINARIES

Let $\ell(x;\xi)$ be the per-sample loss for $x \in \mathbb{R}^n$, and $\mathcal{D}$ the data distribution, with training set $S \sim \mathcal{D}^{|S|}$. Define the population, training, and mini-batch objectives:

$$f_{\mathcal{D}}(x) = \mathbb{E}_{\xi\sim\mathcal{D}}[\ell(x;\xi)], \qquad f(x) = \frac{1}{|S|}\sum_{\xi\in S}\ell(x;\xi), \qquad f_B(x) = \frac{1}{|B|}\sum_{\xi\in B}\ell(x;\xi). \tag{2}$$

Adam preconditioner. We consider Adam and its variants (e.g., AdamW) as the base optimizer. Adam maintains a diagonal preconditioner; at iteration t,

$$D_t = \mathrm{Diag}\Big(\sqrt{\hat{v}_t} + c\Big) \succ 0, \tag{3}$$

where $\hat{v}_t$ is the bias-corrected second moment estimate and $c > 0$ is a small constant. The full Adam(W) update, including momentum and bias correction, is given in Appendix A.2.

Preconditioned stochastic gradient descent. Rather than analyzing Adam directly, we work with a preconditioned SGD that simplifies it, using a positive-definite preconditioner $P_t \succ 0$:

$$x_{t+1} = x_t - \eta_t P_t^{-1} g_t, \tag{4}$$

where $g_t = \nabla f_B(x_t)$ is a mini-batch gradient and P_t is treated as given; P_t need not be diagonal. We defer momentum and weight decay, which extend equation 4 to Adam(W), to Appendix A.

SAM. For a radius $\rho > 0$, SAM (Foret et al., 2021) minimizes the worst-case loss over a Euclidean ball, as in equation 1. Approximating the maximization to first order with the mini-batch gradient g_t gives the perturbation

$$\epsilon_t = \rho\frac{g_t}{\|g_t\|_2}. \tag{5}$$

The perturbed gradient $g_t^+ = \nabla f_B(x_t + \epsilon_t)$ is then passed to the base optimizer in place of g_t.

Preconditioned SAM. Preconditioned SAM (PSAM) (Singh et al., 2025) was proposed to reconcile the geometries of SAM and Adam. PSAM replaces the SAM perturbation with

$$\epsilon_t = \rho\frac{P_t^{-1}g_t}{\|P_t^{-1}g_t\|_2}. \tag{6}$$

Our analysis assumes that the perturbation and the update share the same P_t. With Adam, however, we take $P_t = D_{t-1}$, as the update's D_t is not yet available when the perturbation is computed; we revisit this mismatch in Section 6.1. The full PSAM + Adam(W) update is given in Appendix A.2.

4 PSAM IN SCALED GEOMETRY

We analyze the PSAM perturbation and update under a preconditioner-scaled coordinate change. The key insight is that, at iteration t, a preconditioned PSAM update in x-space is exactly an SGD update with standard SAM in the P_t-scaled space.

4.1 PSAM PERTURBATION IN SCALED GEOMETRY

We interpret the PSAM perturbation equation 6 as the first-order maximizer over an ellipsoidal neighborhood. The perturbation $\epsilon_t \propto P_t^{-1}g_t$ satisfies $\epsilon_t^\top P_t \epsilon_t = \widetilde{\rho}_t^2$, so it lies on the boundary of the ellipsoidal neighborhood $\mathcal{N}_t = \{\epsilon : \epsilon^\top P_t \epsilon \le \widetilde{\rho}_t^2\}$, with effective radius

$$\widetilde{\rho}_t = \rho\frac{\left\|P_t^{-1/2}g_t\right\|_2}{\|P_t^{-1}g_t\|_2}. \tag{7}$$

By the generalized Cauchy–Schwarz inequality, ϵ_t maximizes $g_t^\top\epsilon$ over $\mathcal{N}_t$, giving a stochastic first-order approximation to $f^{\mathrm{PSAM}}(x;\widetilde{\rho}_t)$, where, for $r > 0$,

$$f^{\mathrm{PSAM}}(x;r) := \max_{\epsilon^\top P_t\epsilon\le r^2} f(x+\epsilon). \tag{8}$$

Under the coordinate change

$$\widetilde{x} = P_t^{1/2} x, \qquad \widetilde{\epsilon} = P_t^{1/2} \epsilon, \qquad \widetilde{f} := f \circ P_t^{-1/2}, \qquad \widetilde{f}_B := f_B \circ P_t^{-1/2}, \tag{9}$$

$\mathcal{N}_t$ maps to the Euclidean ball of radius $\widetilde{\rho}_t$, so that

$$f^{\mathrm{PSAM}}(x; \widetilde{\rho}_t) = \max_{\|\widetilde{\epsilon}\|_2 \le \widetilde{\rho}_t} \widetilde{f}(\widetilde{x} + \widetilde{\epsilon}). \tag{10}$$

In $\widetilde{x}$-space, the PSAM perturbation $\widetilde{\epsilon}_t$ is thus the standard SAM perturbation for $\widetilde{f}_B$ at radius $\widetilde{\rho}_t$, and the PSAM objective $f^{\mathrm{PSAM}}(x; \widetilde{\rho}_t)$ is exactly the standard SAM objective for $\widetilde{f}$.

4.2 PSAM UPDATE IN SCALED GEOMETRY

Combining the preconditioned update equation 4 with the PSAM perturbation gives

$$x_{t+1} = x_t - \eta_t P_t^{-1} g_t^+. \tag{11}$$

The following proposition shows that this update is exactly a SAM update in the scaled coordinates.

Proposition 4.1 (PSAM as scaled-coordinate SAM)**.** *Let $\widetilde{\cdot}$ denote the P_t-scaled counterpart at iteration t, as defined in the previous subsection. The update equation 11 becomes*

$$\widetilde{x}_{t+1} = \widetilde{x}_t - \eta_t \nabla_{\widetilde{x}} \widetilde{f}_B(\widetilde{x}_t + \widetilde{\epsilon}_t), \tag{12}$$

which is exactly an SGD step with standard SAM in $\widetilde{x}$-space with radius $\widetilde{\rho}_t$.

Proof. Applying $P_t^{1/2}$ to equation 11,

$$\widetilde{x}_{t+1} = P_t^{1/2} x_{t+1} = \widetilde{x}_t - \eta_t P_t^{-1/2} \nabla f_B(x_t + \epsilon_t), \tag{13}$$

and $P_t^{-1/2} \nabla f_B(x_t + \epsilon_t) = \nabla_{\widetilde{x}} \widetilde{f}_B(\widetilde{x}_t + \widetilde{\epsilon}_t)$ by the chain rule. □

The extension to momentum and weight decay is given in Appendix A.3.

5 THEORY IN SCALED GEOMETRY

Since each PSAM step is exactly an SGD step with standard SAM in its P_t-scaled space (Section 4), we can adapt SAM theory to PSAM. We state stationarity and PAC-Bayes generalization guarantees, then analyze a one-step descent property and admissible radius.

5.1 STATIONARITY

In the nonconvex stochastic setting, we measure progress through the average gradient norm, bounding it for the PSAM update equation 11 in both the preconditioned and Euclidean norms. We first list the assumptions used in the analysis. Throughout, g_t denotes the mini-batch gradient.

Assumption 5.1 (Smooth training objective)**.** *f is L-smooth and bounded below by $f_\star > -\infty$.*

Assumption 5.2 (Local preconditioned smoothness)**.** *There exists $\widetilde{L} > 0$ such that, for all t almost surely,*

$$\left\| P_t^{-1/2} (g_t^+ - g_t) \right\|_2 \le \widetilde{L} \widetilde{\rho}_t. \tag{14}$$

This is a smoothness condition in the scaled space, imposed locally along the perturbation segment rather than globally, which makes it more plausible.

Assumption 5.3 (Preconditioned stochastic oracle)**.** *Let $\mathbb{E}_t$ denote the conditional expectation given x_t and P_t. Then $\mathbb{E}_t[g_t] = \nabla f(x_t)$ and $\mathbb{E}_t[\|P_t^{-1} g_t^+\|_2^2] \le G^2$.*

The bound is on the actual update direction $P_t^{-1} g_t^+$, which is self-normalized for Adam-type preconditioners, where P_t tracks the second moment of past gradients.

Assumption 5.4 (Bounded preconditioner). *$P_t \preceq d_{\max} I$ for all t.*

Building on Theorem 4.6 of Si & Yun (2023), which establishes gradient norm bounds for standard SAM with SGD, one might expect a similar result for PSAM in the scaled space (Proposition 4.1). The smoothness assumption, however, becomes problematic under this transfer: global smoothness is hard to justify in the scaled space, where small preconditioner coordinates make $P_t^{-1/2}$ large. We therefore use the local form of Assumption 5.2, which requires modifying parts of the argument.

Theorem 5.5 (Stationarity in preconditioned norm). *Under Assumptions 5.1–5.3, let $\Delta = f(x_0) - f_\star$. For the PSAM update equation 11 run for T steps with fixed step size $\eta_t = \eta > 0$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\| P_t^{-1/2} \nabla f(x_t) \right\|_2^2\right] \le \frac{2\Delta}{\eta T} + \widetilde{L}^2 \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\widetilde{\rho}_t^2\right] + L\eta G^2. \tag{15}$$

The proof is in Appendix B.1. The bound matches that of standard SAM with ρ replaced by the iteration-dependent effective radius $\widetilde{\rho}_t$. We next convert this to the Euclidean norm.

Corollary 5.6 (Stationarity in Euclidean norm). *In the setting of Theorem 5.5, further assume Assumption 5.4 and set $\eta = \sqrt{2\Delta/(LG^2T)}$. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] \le d_{\max}\left(2\sqrt{\frac{2L\Delta G^2}{T}} + d_{\max}\widetilde{L}^2\rho^2\right). \tag{16}$$

The proof is in Appendix B.2. Assumption 5.4 supplies the bounds driving this conversion, such as $\widetilde{\rho}_t^2 \le d_{\max}\rho^2$. This gives a bound of the form $\mathcal{O}(1/\sqrt{T}) + \mathcal{O}(\rho^2)$, matching the rate of standard SAM. Together, Theorem 5.5 and Corollary 5.6 show that PSAM enjoys a stationarity guarantee comparable to that of standard SAM (Si & Yun, 2023).

5.2 Generalization

We give a PAC-Bayes generalization bound for PSAM at a fixed iteration, treating $P \succ 0$ and $\widetilde{\rho} > 0$ as fixed. By equation 10, the PSAM objective is the SAM objective in the scaled space, so we adapt the assumption and bound of Foret et al. (2021) to this space.

Assumption 5.7 (Preconditioned Gaussian smoothing). *At a point $x \in \mathbb{R}^n$, assume that $f_{\mathcal{D}}(x) \le \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\widetilde{\rho}^2 P^{-1})}[f_{\mathcal{D}}(x+\epsilon)]$.*

Corollary 5.8 (PAC-Bayes generalization bound for PSAM). *Assume that the per-sample loss is bounded in $[0,1]$. Fix $\delta \in (0,1)$ and a point $x \in \mathbb{R}^n$ satisfying Assumption 5.7. Then, with probability at least $1-\delta$ over the training set S,*

$$f_{\mathcal{D}}(x) \le f^{\mathrm{PSAM}}(x;\widetilde{\rho}) + \mathrm{comp}\left(\frac{x^\top P x}{\widetilde{\rho}^2}, |S|, \delta\right), \tag{17}$$

where comp *is increasing in its first argument; see Theorem 2 of Foret et al. (2021).*

The proof is in Appendix B.3. The scaled term $x^\top P x/\widetilde{\rho}^2$ replaces the Euclidean term $\|x\|_2^2/\rho^2$ in standard SAM. Under this fixed-metric interpretation, Corollary 5.8 provides a PAC-Bayes guarantee for PSAM comparable to that of standard SAM. In practice, P_t and $\widetilde{\rho}_t$ are data-dependent, so this result does not directly extend to the full training trajectory.

5.3 Descent property and admissible radius

This subsection studies under what conditions a SAM update decreases the loss, and transfers this to PSAM through the scaled space. We further show that this property can break when the two geometries are mismatched. We then give a qualitative account of why PSAM can use a wider range of radii. Here we use *quasiconvexity*, a weakened form of convexity; further discussion is in Appendix B.4. We begin by showing that quasiconvexity yields the descent property for SAM.

Theorem 5.9 (Descent property of standard SAM). *Fix an iteration t with mini-batch B, and let $x_t^+ = x_t + \epsilon_t$ with the SAM perturbation equation 5. Suppose f_B is L_B-smooth and quasiconvex on an open convex set $\mathcal{Q}$, with $x_t, x_t^+ \in \mathcal{Q}$ and $g_t, g_t^+ \neq 0$. Then $\langle g_t, g_t^+\rangle > 0$, and for all sufficiently small $\eta_t > 0$ the update $x_{t+1} = x_t - \eta_t g_t^+$ satisfies $f_B(x_{t+1}) < f_B(x_t)$ and $x_{t+1} \in \mathcal{Q}$.*

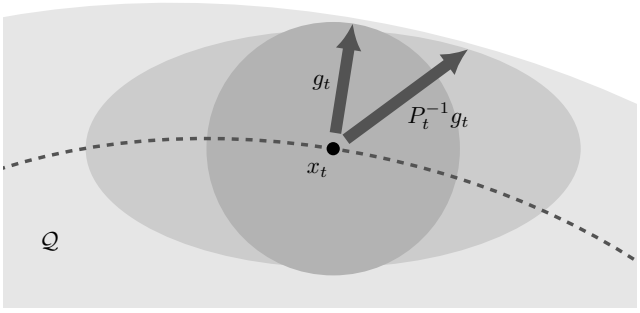


Figure 2: Why PSAM can admit larger radii. From x_t, SAM perturbs along g_t and PSAM along $P_t^{-1}g_t$, each reaching the boundary of the convex sublevel set $\mathcal{Q}$. Because the PSAM direction is tilted away from the steepest ascent g_t, it travels farther before leaving $\mathcal{Q}$, so PSAM tolerates a larger radius. The dashed curve is the level set through x_t. The inner circle and ellipse are the SAM and PSAM neighborhoods, respectively.

The proof is in Appendix B.5. Quasiconvexity aligns g_t and g_t^+ through a supporting hyperplane argument (Lemma B.4).

The same holds for PSAM: since it is standard SAM in the scaled space (Proposition 4.1) and quasiconvexity is preserved there (Lemma B.3), it can be proved as in Theorem 5.9.

Corollary 5.10 (Descent property of PSAM)**.** *Fix an iteration t with mini-batch B, and let $x_t^+ = x_t + \epsilon_t$ with the PSAM perturbation equation 6. Suppose f_B is L_B-smooth and quasiconvex on an open convex set $\mathcal{Q}$, with $x_t, x_t^+ \in \mathcal{Q}$ and $g_t, g_t^+ \neq 0$. Then $g_t^\top P_t^{-1} g_t^+ > 0$, and for all sufficiently small $\eta_t > 0$ the update $x_{t+1} = x_t - \eta_t P_t^{-1} g_t^+$ satisfies $f_B(x_{t+1}) < f_B(x_t)$ and $x_{t+1} \in \mathcal{Q}$.*

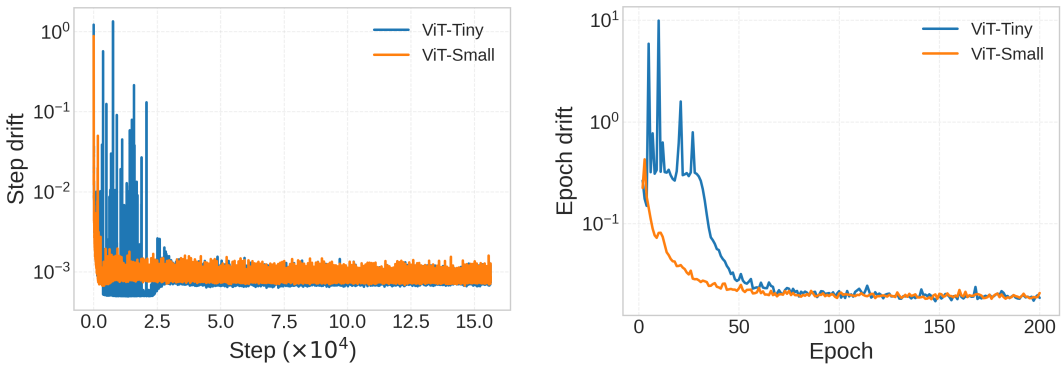
The proof is in Appendix B.6. The gradient alignment comes from the scaled space, while the descent property relies on the smoothness of f_B in the original space, not in the scaled one.

Mismatched geometries lose the descent property. The preconditioned update equation 4 and a standard SAM perturbation belong to different geometries. Quasiconvexity still gives $\langle g_t, g_t^+ \rangle > 0$, but descent now requires $\langle g_t, P_t^{-1} g_t^+ \rangle > 0$, which does not follow: a quadratic counterexample is given in Appendix B.7. Such a step may still decrease the loss, but quasiconvexity no longer guarantees it. By matching the perturbation to the scaled geometry, PSAM restores this guarantee, which makes it well suited to Adam-type optimizers.

Larger admissible radius. The descent property suggests which radii are safe to use. For a given perturbation rule, we call ρ *admissible* if descent is guaranteed for every radius up to ρ. By the results above, a sufficient condition is that the perturbation at radius ρ lies in $\mathcal{Q}$: since $x_t \in \mathcal{Q}$ and $\mathcal{Q}$ is convex, all smaller perturbations along the same direction also remain in $\mathcal{Q}$.

To discuss this further, suppose that $\mathcal{Q}$ is a strict sublevel set of f_B. To first order, f_B increases fastest along the gradient direction g_t. SAM perturbs exactly along this direction, whereas PSAM perturbs along $P_t^{-1} g_t$, which is generally not aligned with g_t. Thus, for the same Euclidean radius, PSAM can increase f_B more slowly and remain in $\mathcal{Q}$ up to a larger radius (Figure 2). This suggests that PSAM can have a larger admissible radius than SAM, especially when $P_t^{-1} g_t$ points away from g_t, which we examine in Section 7.

We note that this argument is heuristic, as it relies on a first-order expansion in ρ. A practical question is how plausible $x_t \in \mathcal{Q}$ is in deep learning. Loss landscape visualizations often show a convex-looking level set in low-loss regions (Li et al., 2018), as we also observe (Appendix D.4). Starting SAM from low loss therefore makes the assumption more plausible, providing one motivation for a warm start.



(a) Step-wise drift.

(b) Epoch-wise drift.

Figure 3: The preconditioner drift is large early in training and later settles to a small value. For ViT-Tiny and ViT-Small on CIFAR-100, the drift is measured (a) step-wise and (b) epoch-wise during a single PSAM training run ($\rho = 1$). The step-wise curve uses a 10-step rolling mean.

6 Warm-started PSAM

Since our theoretical assumptions depend on P_t, preconditioner stability is important for the analysis. In this section, we empirically measure the drift of the AdamW preconditioner during PSAM training, and discuss why PSAM requires a warm start, particularly at large radii.

6.1 Preconditioner stability

We first quantify how rapidly the AdamW preconditioner changes during training. Let D_k denote the preconditioner at time k. We define the preconditioner drift as

$$\text{Drift}_k = \frac{\|D_k - D_{k-1}\|_F}{\|D_{k-1}\|_F}.$$

When k indexes training steps, this gives the step-wise drift; when k indexes epoch endpoints, it gives the epoch-wise drift. The step-wise drift measures the mismatch between the preconditioners used for perturbation and update in PSAM, while the epoch-wise drift measures longer-term stability needed for our theoretical assumptions to remain plausible.

We assess preconditioner stability directly through PSAM experiments. We train ViT-Tiny and ViT-Small from scratch on CIFAR-100 using PSAM with $\rho = 1$, reporting a single run. The results are shown in Figure 3. Both the step-wise and epoch-wise drift show that the preconditioner is highly unstable early in training and stabilizes later. The other quantities entering the analysis of Section 5 show a similar pattern in the same runs, with substantial early fluctuations followed by more stable behavior later in training (Appendix C).

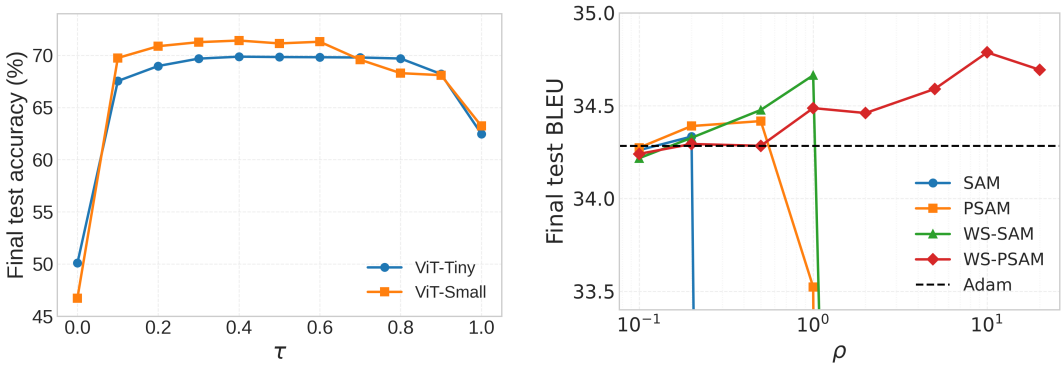
6.2 Warm-start method

This instability motivates delaying the PSAM update until the preconditioner has settled. Given total iterations T and a ratio $\tau \in [0, 1]$, a warm-started method feeds the base optimizer g_t for $t < \tau T$ and the perturbed gradient g_t^+ for $t \geq \tau T$, which also skips the second gradient evaluation of SAM-type updates during the first τT iterations. We denote the warm-started variants of SAM and PSAM by WS-SAM and WS-PSAM, respectively, with WS-SAM serving as a baseline.

Why PSAM needs a warm start. Warm-starting is especially effective for PSAM, for two reasons. First, the PSAM perturbation depends on the unstable early preconditioner, whereas the SAM perturbation does not. Second, a warm start lets training tolerate larger radii, because it avoids the hallucinated minimizers that SAM can fall into at large ρ (Park et al., 2026). This effect combines with PSAM's own tolerance for a wider range of ρ (Section 5.3): PSAM can use large radii more stably than SAM, and thus benefit more from the generalization gains they bring.

Table 1: WS-PSAM achieves the best accuracy on both ViT models. CIFAR-100 final test accuracy (%) for ViT-Tiny and ViT-Small, over three seeds (mean ± std). For each SAM-type method, we report the value at its best-performing ρ. WS-SAM and WS-PSAM use a warm-start ratio $\tau = 0.4$.

Model	AdamW	SAM	PSAM	WS-SAM	WS-PSAM
ViT-Tiny	62.47 ± 0.51	64.48 ± 1.57	64.14 ± 0.72	68.99 ± 0.24	$\mathbf{69.87 \pm 0.18}$
ViT-Small	63.22 ± 0.28	66.74 ± 0.11	66.78 ± 0.45	68.36 ± 2.09	$\mathbf{71.42 \pm 0.28}$



(a) Warm-start ratio sweep.

(b) IWSLT14 De-En.

Figure 4: (a) Final test accuracy of WS-PSAM ($\rho = 1$) on CIFAR-100 as the warm-start ratio τ varies; $\tau = 0$ corresponds to PSAM and $\tau = 1$ to AdamW. Any warm start improves on PSAM. (b) Final test BLEU on IWSLT14 De-En versus ρ, with the Adam baseline. WS-PSAM remains effective at radii where the other methods collapse. All results are averaged over three seeds.

7 Experiments

We evaluate WS-PSAM against SAM, PSAM, and WS-SAM on image classification, machine translation, and language modeling, sweeping the radius ρ in each setting. In particular, we verify that WS-PSAM performs well at relatively large ρ across several deep learning tasks with Adam-type base optimizers.

7.1 Image classification

We train ViT-Tiny and ViT-Small (Dosovitskiy et al., 2021) from scratch on CIFAR-100 (Krizhevsky, 2009) with AdamW as the base optimizer and a constant learning rate, reporting the final test accuracy as a function of ρ. WS-SAM and WS-PSAM use a warm-start ratio $\tau = 0.4$. The results are shown in Table 1; see Appendix D.1 for the experimental setup and full results. For both models, WS-PSAM achieves the best performance.

We further study the effect of the warm-start ratio τ (Figure 4a; see Appendix D.1 for detailed results). For both models, $\tau = 0.4$ gives the best performance. An overly large τ degrades it, as the SAM-type update has too little time to act. This interacts with the learning-rate schedule: if switched on too late, it acts only after the learning rate has decayed, so the appropriate τ shifts. Indeed, our later experiments use a cosine learning-rate scheduler, under which $\tau = 0.2$ works well.

7.2 Machine translation

We train a Transformer (Vaswani et al., 2017) on IWSLT14 German-English (Cettolo et al., 2014) with Adam as the base optimizer, reporting the final test BLEU as a function of ρ. WS-SAM and WS-PSAM use a warm-start ratio $\tau = 0.2$. The results are shown in Figure 4b; see Appendix D.2 for the experimental setup and full results. WS-PSAM performs best, and unlike the other methods it remains effective at relatively large ρ.

Table 2: Cosine similarity between g_t and $P_t^{-1}g_t$ at the final training step for the best WS-PSAM ρ, over three seeds (mean $\pm$ std). Best ρ is reported as WS-SAM / WS-PSAM.

	ViT-Tiny	ViT-Small	Transformer	GPT-2 (30M)	GPT-2 (124M)
Best ρ	$1/1$	$1/1$	$1/10$	$0.2/5$	$1/5$
$\cos(g_t, P_t^{-1}g_t)$	0.501 ± 0.050	0.402 ± 0.028	0.058 ± 0.011	0.040 ± 0.002	0.060 ± 0.017

7.3 Language modeling

We train GPT-2 (30M) on WikiText-2 and GPT-2 (124M) on WikiText-103 (Merity et al., 2017) with AdamW as the base optimizer, reporting the final validation perplexity as a function of ρ. WS-SAM and WS-PSAM use a warm-start ratio $\tau = 0.2$. The results are shown in Figure 1; see Appendix D.3 for the experimental setup and full results. For both models, WS-PSAM achieves the lowest perplexity, and unlike the other methods it remains effective at relatively large ρ.

7.4 Large-radius behavior

Across tasks, the performance of SAM and PSAM deteriorates at large values of ρ, with collapse in several settings. Warm-starting alleviates this sensitivity: WS-SAM and WS-PSAM remain effective over wider ranges of ρ than SAM and PSAM, respectively. The effect is particularly pronounced for PSAM, as several large-radius configurations that perform poorly or fail under PSAM become the best-performing configurations under WS-PSAM.

We now compare the two warm-started methods, whose radii are on the same Euclidean scale and thus directly comparable. On both image-classification models they peak at the same value, $\rho = 1$. In the NLP tasks, however, WS-PSAM remains effective over a substantially wider range of radii, while WS-SAM deteriorates or collapses earlier. Past its best ρ, WS-PSAM falls below WS-SAM only on the vision models; we discuss this in Appendix D.1.

The admissible radius of Section 5.3 is not directly measurable, but a wider admissible range may allow PSAM to benefit from larger radii. We therefore examine whether the cosine similarity between g_t and $P_t^{-1}g_t$ is related to the gap in best ρ. As shown in Table 2, the similarity is relatively high for the vision models and close to zero for the NLP tasks, matching where the gap appears. This correspondence supports the geometric explanation: when the perturbation alignment is low, PSAM can admit a larger radius.

The loss-plane visualization for GPT-2 (30M) illustrates this geometry through elongated, convex-like level sets (Appendix D.4). Along the PSAM direction, a much larger perturbation produces a loss increase comparable to that along the SAM direction.

8 Conclusion

We have shown that preconditioned SAM can exploit large radii and explained why. First, PSAM can be interpreted as SAM in a scaled geometry, which yields SAM-like guarantees. Since SAM's perturbation is mismatched with Adam-type updates, PSAM can thus outperform SAM with such optimizers, and with a warm start, it did so in all our experiments. Moreover, the best-performing radius of WS-PSAM was up to $25\times$ that of WS-SAM, a gap we analyzed through the admissible radius. This is fundamentally because the PSAM perturbation can differ greatly in direction from SAM's steepest-ascent perturbation. These results provide theoretical guarantees and a practical recipe for PSAM, suggesting that it is promising in its own right.

Beyond Adam-type optimizers, our PSAM analysis may apply to other preconditioned methods, since Proposition 4.1 holds for any positive-definite preconditioner. Warm-starting also consistently improved standard SAM, suggesting that it is useful beyond the failure mode studied by Park et al. (2026). The abrupt switch at τT, however, causes a jump in the training loss (Figure 6), which scheduling ρ to apply the perturbation smoothly may alleviate. Finally, similar trends in both GPT-2 models hint that they persist in larger models; experiments at that scale, as well as a comprehensive comparison with other SAM variants, are left for future work.

AI USE STATEMENT

In this work, we used generative AI tools to help formulate mathematical claims, to assist in the writing of proofs, to implement methods, and to assist with translation. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, to provide critical ingredients for proving mathematical claims, to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, or to interpret results; generating synthetic data sets, cleaning and reformatting datasets, and qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures, to create or edit software code, to draft parts of the paper, to summarize or analyze existing literature, to discover research topics or identify gaps, for brainstorming, to source and search for information, to edit the paper for readability, to identify relevant literature, to format references, and to propose a title or keywords.

We have reviewed all AI-assisted work. We made extensive use of AI coding tools to implement all methods and experiments; the authors reviewed all of the generated code and checked the resulting outputs for correctness. All mathematical claims and proofs were verified by the authors. Literature identified or summarized with AI assistance was checked against the original sources, and AI-drafted or AI-edited text was revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The assumptions behind our theoretical results are stated in Section 5, and complete proofs are given in Appendices A and B, with Proposition 4.1 proved in the main text; Appendix C examines these assumptions empirically. Appendix D describes the setup of every experiment, including models, optimizers, hyperparameters, radius grids, warm-start ratios, and seeds, and reports the full results (Tables 3–8). All datasets are publicly available, and their processing, which follows standard pipelines, is also described in Appendix D. An anonymized implementation of PSAM for the GPT-2 (30M) experiments is available at `https://anonymous.4open.science/r/PSAM-Can-Exploit-Large-Perturbation-Radii-33CC` and is also provided as supplementary material.

REFERENCES

M. Andriushchenko and N. Flammarion. Towards understanding sharpness-aware minimization. *International Conference on Machine Learning*, 2022.

D. Bahri, H. Mobahi, and Y. Tay. Sharpness-aware minimization improves language model generalization. *Association for Computational Linguistics*, 2022.

P. L. Bartlett, P. M. Long, and O. 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.

M. Becker, F. Altrock, and B. Risse. Momentum-SAM: Sharpness aware minimization without computational overhead. *Neural Information Processing Systems*, 2025.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico. Report on the 11th IWSLT evaluation campaign. *International Workshop on Spoken Language Translation*, 2014.

X. Chen, C.-J. Hsieh, and B. Gong. When vision transformers outperform ResNets without pretraining or strong data augmentations. *International Conference on Learning Representations*, 2022.

Z. Chen, J. Zhang, Y. Kou, X. Chen, C.-J. Hsieh, and Q. Gu. Why does sharpness-aware minimization generalize better than SGD? *Neural Information Processing Systems*, 2023.

Y. Dai, K. Ahn, and S. Sra. The crucial role of normalization in sharpness-aware minimization. *Neural Information Processing Systems*, 2023.

L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. *International Conference on Machine Learning*, 2017.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *International Conference on Learning Representations*, 2021.

S. Hochreiter and J. Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

W. Jiang, H. Yang, Y. Zhang, and J. Kwok. An adaptive policy to employ sharpness-aware minimization. *International Conference on Learning Representations*, 2023.

J. Kaddour, L. Liu, R. Silva, and M. J. Kusner. When do flat minima optimizers work? *Neural Information Processing Systems*, 2022.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *International Conference on Learning Representations*, 2017.

P. Khanh, H.-C. Luong, B. Mordukhovich, and D. Tran. Fundamental convergence analysis of sharpness-aware minimization. *Neural Information Processing Systems*, 2024.

H. Kim, J. Park, Y. Choi, and J. Lee. Stability analysis of sharpness-aware minimization. *International Conference on Machine Learning*, 2026.

M. Kim, D. Li, S. X. Hu, and T. M. Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. *International Conference on Machine Learning*, 2022.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

A. Krizhevsky. Learning multiple layers of features from tiny images. *Technical report*, 2009.

J. Kwon, J. Kim, H. Park, and I. K. Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. *International Conference on Machine Learning*, 2021.

B. Li and G. B. Giannakis. Enhancing sharpness-aware optimization through variance suppression. *Neural Information Processing Systems*, 2023.

H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets. *Neural Information Processing Systems*, 2018.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

H. Luo, M. Harandi, D. Phung, and T. Le. Unveiling m-sharpness through the structure of stochastic gradient noise. *Neural Information Processing Systems*, 2025.

S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models. *International Conference on Learning Representations*, 2017.

T. Möllenhoff and M. E. Khan. SAM as an optimal relaxation of Bayes. *International Conference on Learning Representations*, 2023.

C. Moon, D. Si, and C. Yun. Minor first, major last: A depth-induced implicit bias of sharpness-aware minimization. *International Conference on Learning Representations*, 2026.

B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. *Neural Information Processing Systems*, 2017.

D. Oikonomou and N. Loizou. Sharpness-aware minimization: General analysis and improved rates. *International Conference on Learning Representations*, 2025.

C. Park, U. Jang, E. K. Ryu, and I. Yang. Sharpness-aware minimization can hallucinate minimizers. *International Conference on Machine Learning*, 2026.

D. Si and C. Yun. Practical sharpness-aware minimization cannot converge all the way to optima. *Neural Information Processing Systems*, 2023.

S. P. Singh, H. Mobahi, A. Agarwala, and Y. Dauphin. Avoiding spurious sharpness minimization broadens applicability of SAM. *International Conference on Machine Learning*, 2025.

H. Sun, L. Shen, Q. Zhong, L. Ding, S. Chen, J. Sun, J. Li, G. Sun, and D. Tao. AdaSAM: Boosting sharpness-aware minimization with adaptive learning rate and momentum for training deep neural networks. *Neural Networks*, 169:506–519, 2024.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Neural Information Processing Systems*, 2017.

K. Wen, T. Ma, and Z. Li. How does sharpness-aware minimization minimize sharpness? *International Conference on Learning Representations*, 2023.

J. Yun and E. Yang. Riemannian SAM: Sharpness-aware minimization on Riemannian manifolds. *Neural Information Processing Systems*, 2023.

Y. Zhang, B. Li, and G. B. Giannakis. Preconditioned sharpness-aware minimization: Unifying analysis and a novel learning algorithm. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Z. Zhou, M. Wang, Y. Mao, B. Li, and J. Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. *International Conference on Learning Representations*, 2025.

A EXTENSION TO ADAM(W)

The preconditioned SGD equation 4 omits momentum and weight decay. We restore them to recover Adam(W) (Kingma & Ba, 2015; Loshchilov & Hutter, 2019), write out the full PSAM + Adam(W) update, and extend Proposition 4.1 to momentum and weight decay.

A.1 EXTENSION OF PRECONDITIONED SGD TO ADAM(W)

Adding momentum and weight decay to preconditioned SGD equation 4 gives, with a diagonal preconditioner $P_t \succ 0$, step size $\eta_t > 0$, momentum coefficient $\beta_1 \in [0, 1)$, and weight decay $\lambda \geq 0$,

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \tag{18}$$

$$x_{t+1} = x_t - \eta_t P_t^{-1} m_t - \eta_t \lambda x_t. \tag{19}$$

Maintaining the second moment $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$ with $\beta_2 \in [0, 1)$ and choosing

$$P_t = (1 - \beta_1^t)\ \mathrm{Diag}\big(\sqrt{v_t/(1 - \beta_2^t)} + c\big) \tag{20}$$

gives $\eta_t P_t^{-1} m_t = \eta_t\, \hat{m}_t/(\sqrt{\hat{v}_t} + c)$, with bias-corrected moments $\hat{m}_t = m_t/(1 - \beta_1^t)$ and $\hat{v}_t = v_t/(1 - \beta_2^t)$, so

$$x_{t+1} = x_t - \eta_t\, \hat{m}_t/(\sqrt{\hat{v}_t} + c) - \eta_t \lambda x_t, \tag{21}$$

which is AdamW; $\lambda = 0$ recovers Adam. Adam(W) is thus an instance of preconditioned SGD whose preconditioner is defined by the gradient second moment.

A.2 FULL PSAM + ADAM(W) UPDATE

PSAM feeds the perturbed gradient g_t^+ into this optimizer in place of g_t. The full PSAM + Adam(W) update at iteration t is

$$\begin{aligned}
\epsilon_t &= \rho \frac{P_t^{-1} g_t}{\left\| P_t^{-1} g_t \right\|_2}, \\
g_t^+ &= \nabla f_B(x_t + \epsilon_t), \\
m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t^+, \\
v_t &= \beta_2 v_{t-1} + (1 - \beta_2)(g_t^+)^2, \\
\hat{m}_t &= m_t/(1 - \beta_1^t), \qquad \hat{v}_t = v_t/(1 - \beta_2^t), \\
D_t &= \mathrm{Diag}(\sqrt{\hat{v}_t} + c), \\
x_{t+1} &= x_t - \eta_t D_t^{-1} \hat{m}_t - \eta_t \lambda x_t,
\end{aligned}$$

where $\rho, \eta_t > 0$ and $c > 0$ is a small stability constant. Setting $\lambda = 0$ gives the PSAM + Adam case. Here P_t is used only for the perturbation; the base optimizer runs on D_t.

Viewing Adam(W) as preconditioned SGD as before gives $P_t = (1 - \beta_1^t) D_t$. Since the perturbation is formed before D_t is available, practical PSAM instead uses $P_t = D_{t-1}$. The scalar factor cancels in the normalized perturbation, so we drop the bias correction.

A.3 EQUIVALENCE FOR EXTENDED PRECONDITIONED SGD

We extend Proposition 4.1 to preconditioned SGD with momentum and weight decay.

Proposition A.1 (PSAM as scaled-coordinate SAM). *Let $\widetilde{\cdot}$ denote the P_t-scaled counterpart at iteration t, as in Proposition 4.1. The update in equations 18–19, with g_t replaced by the perturbed gradient g_t^+, becomes*

$$\widetilde{m}_t = \beta_1 \widetilde{m}_{t-1} + (1 - \beta_1) \widetilde{g}_t^+, \tag{22}$$

$$\widetilde{x}_{t+1} = \widetilde{x}_t - \eta_t \widetilde{m}_t - \eta_t \lambda \widetilde{x}_t, \tag{23}$$

which is an SGD step with momentum, weight decay, and standard SAM in $\widetilde{x}$-space with radius $\widetilde{\rho}_t$.

Proof. Applying $P_t^{-1/2}$ to the momentum recurrence,

$$\widetilde{m}_t = P_t^{-1/2} m_t = \beta_1 P_t^{-1/2} m_{t-1} + (1-\beta_1) P_t^{-1/2} g_t^+ = \beta_1 \widetilde{m}_{t-1} + (1-\beta_1)\widetilde{g}_t^+, \tag{24}$$

where $\widetilde{g}_t^+ = P_t^{-1/2} g_t^+ = \nabla_{\widetilde{x}} \widetilde{f}_B(\widetilde{x}_t + \widetilde{\epsilon}_t)$ by the chain rule (cf. Proposition 4.1). Applying $P_t^{1/2}$ to the parameter update,

$$\widetilde{x}_{t+1} = P_t^{1/2} x_{t+1} = \widetilde{x}_t - \eta_t P_t^{-1/2} m_t - \eta_t \lambda \widetilde{x}_t = \widetilde{x}_t - \eta_t \widetilde{m}_t - \eta_t \lambda \widetilde{x}_t. \tag{25}$$

□

Thus, with the current preconditioner D_t used for the perturbation, the same per-step equivalence extends to Adam(W).

B Details for Section 5

B.1 Proof of Theorem 5.5

Let $q_t = \left\| P_t^{-1/2} \nabla f(x_t) \right\|_2^2$. By L-smoothness (Assumption 5.1) and equation 11,

$$f(x_{t+1}) \le f(x_t) - \eta \left\langle \nabla f(x_t), P_t^{-1} g_t^+ \right\rangle + \frac{L\eta^2}{2} \left\| P_t^{-1} g_t^+ \right\|_2^2. \tag{26}$$

Taking conditional expectation and using Assumption 5.3,

$$\mathbb{E}_t[f(x_{t+1})] \le f(x_t) - \eta \left\langle \nabla f(x_t), P_t^{-1} \mathbb{E}_t[g_t^+] \right\rangle + \frac{L\eta^2}{2} G^2. \tag{27}$$

We decompose the inner product using $\mathbb{E}_t[g_t] = \nabla f(x_t)$:

$$\left\langle \nabla f(x_t), P_t^{-1} \mathbb{E}_t[g_t^+] \right\rangle = q_t + \left\langle \nabla f(x_t), P_t^{-1} \mathbb{E}_t[g_t^+ - g_t] \right\rangle. \tag{28}$$

For the second term, note that $\nabla f(x_t)$ is fixed given x_t, hence constant under $\mathbb{E}_t$. Applying Cauchy–Schwarz to $P_t^{-1/2} \nabla f(x_t)$ and $P_t^{-1/2} \mathbb{E}_t[g_t^+ - g_t]$, then moving the norm inside the expectation by Jensen's inequality and using Assumption 5.2,

$$\left| \left\langle \nabla f(x_t), P_t^{-1} \mathbb{E}_t[g_t^+ - g_t] \right\rangle \right| \le \left\| P_t^{-1/2} \nabla f(x_t) \right\|_2 \cdot \left\| P_t^{-1/2} \mathbb{E}_t[g_t^+ - g_t] \right\|_2 \tag{29}$$

$$\le \left\| P_t^{-1/2} \nabla f(x_t) \right\|_2 \cdot \mathbb{E}_t \left[\left\| P_t^{-1/2} (g_t^+ - g_t) \right\|_2 \right] \tag{30}$$

$$\le \sqrt{q_t} \cdot \widetilde{L}\, \mathbb{E}_t[\widetilde{\rho}_t]. \tag{31}$$

The bound on the second term gives

$$\left\langle \nabla f(x_t), P_t^{-1} \mathbb{E}_t[g_t^+] \right\rangle \ge q_t - \sqrt{q_t}\, \widetilde{L}\, \mathbb{E}_t[\widetilde{\rho}_t]. \tag{32}$$

Applying Young's inequality $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$ with $a = \sqrt{q_t}$ and $b = \widetilde{L}\, \mathbb{E}_t[\widetilde{\rho}_t]$, followed by Jensen's inequality $\mathbb{E}_t[\widetilde{\rho}_t]^2 \le \mathbb{E}_t[\widetilde{\rho}_t^2]$,

$$\left\langle \nabla f(x_t), P_t^{-1} \mathbb{E}_t[g_t^+] \right\rangle \ge \frac{1}{2} q_t - \frac{1}{2} \widetilde{L}^2 \mathbb{E}_t[\widetilde{\rho}_t^2]. \tag{33}$$

Substituting equation 33 into equation 27, rearranging, and taking total expectation,

$$\frac{\eta}{2} \mathbb{E}[q_t] \le \mathbb{E}[f(x_t) - f(x_{t+1})] + \frac{\eta \widetilde{L}^2}{2} \mathbb{E}[\widetilde{\rho}_t^2] + \frac{L\eta^2 G^2}{2}. \tag{34}$$

Summing over $t = 0, \ldots, T-1$, using $f(x_T) \ge f_\star$, and dividing by $\eta T/2$ gives equation 15. □

B.2 Proof of Corollary 5.6

Under Assumption 5.4, $P_t^{-1} \succeq d_{\max}^{-1} I$, so

$$\|\nabla f(x_t)\|_2^2 \le d_{\max} \cdot \nabla f(x_t)^\top P_t^{-1} \nabla f(x_t) = d_{\max} \left\| P_t^{-1/2} \nabla f(x_t) \right\|_2^2. \tag{35}$$

Letting $u = P_t^{-1} g_t$,

$$\widetilde{\rho}_t^2 = \rho^2 \frac{g_t^\top P_t^{-1} g_t}{g_t^\top P_t^{-2} g_t} = \rho^2 \frac{u^\top P_t u}{u^\top u} \le \rho^2 \lambda_{\max}(P_t) \le d_{\max} \rho^2, \tag{36}$$

by the Rayleigh quotient bound and $P_t \preceq d_{\max} I$.

Substituting these bounds into Theorem 5.5,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|_2^2 \right] \le d_{\max} \left(\frac{2\Delta}{\eta T} + L\eta G^2 + d_{\max} \widetilde{L}^2 \rho^2 \right). \tag{37}$$

Choosing $\eta = \sqrt{2\Delta/(LG^2T)}$ balances the first and second terms into $2\sqrt{2L\Delta G^2/T}$, giving the result.

B.3 Proof of Corollary 5.8

Fix $P \succ 0$ and $\widetilde{\rho} > 0$. Under the coordinate change equation 9, with P in place of P_t, define $\widetilde{f}_{\mathcal{D}} := f_{\mathcal{D}} \circ P^{-1/2}$. For $\widetilde{\epsilon} \sim \mathcal{N}(0, \widetilde{\rho}^2 I)$, we have $P^{-1/2}\widetilde{\epsilon} \sim \mathcal{N}(0, \widetilde{\rho}^2 P^{-1})$, so Assumption 5.7 becomes

$$\widetilde{f}_{\mathcal{D}}(\widetilde{x}) \le \mathbb{E}_{\widetilde{\epsilon} \sim \mathcal{N}(0, \widetilde{\rho}^2 I)} [\widetilde{f}_{\mathcal{D}}(\widetilde{x} + \widetilde{\epsilon})]. \tag{38}$$

Theorem (Foret et al., 2021, Theorem 2)**.** *Assume that the per-sample loss is bounded in* $[0, 1]$*. Fix* $\delta \in (0, 1)$*. For* $\rho > 0$ *and a point* $x \in \mathbb{R}^n$*, suppose that* $f_{\mathcal{D}}(x) \le \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \rho^2 I)}[f_{\mathcal{D}}(x + \epsilon)]$*. Then, with probability at least* $1 - \delta$ *over the sample* S*,*

$$f_{\mathcal{D}}(x) \le \max_{\|\epsilon\|_2 \le \rho} f(x + \epsilon) + \mathrm{comp}\left(\frac{\|x\|_2^2}{\rho^2}, |S|, \delta \right). \tag{39}$$

The bounded-loss condition is preserved under the coordinate change. Applying the theorem to $\widetilde{f}$ at radius $\widetilde{\rho}$ and at the point $\widetilde{x} = P^{1/2} x$ gives

$$\widetilde{f}_{\mathcal{D}}(\widetilde{x}) \le \max_{\|\widetilde{\epsilon}\|_2 \le \widetilde{\rho}} \widetilde{f}(\widetilde{x} + \widetilde{\epsilon}) + \mathrm{comp}\left(\frac{\|\widetilde{x}\|_2^2}{\widetilde{\rho}^2}, |S|, \delta \right). \tag{40}$$

Since $\widetilde{f}_{\mathcal{D}}(\widetilde{x}) = f_{\mathcal{D}}(x)$, $\max_{\|\widetilde{\epsilon}\|_2 \le \widetilde{\rho}} \widetilde{f}(\widetilde{x} + \widetilde{\epsilon}) = f^{\mathrm{PSAM}}(x; \widetilde{\rho})$ by equation 10, and $\|\widetilde{x}\|_2^2 = x^\top P x$, this is the claimed bound.

B.4 On the quasiconvexity assumption

Here we introduce the notion of quasiconvexity used in Section 5.3.

Definition B.1 (Quasiconvexity)**.** *A function* f *is* quasiconvex *on a convex set* $\mathcal{C}$ *if every sublevel set* $\{x \in \mathcal{C} : f(x) \le \alpha\}$ *is convex, equivalently, if* $f(\lambda x + (1 - \lambda) y) \le \max\{f(x), f(y)\}$ *for all* $x, y \in \mathcal{C}$ *and* $\lambda \in [0, 1]$*.*

Note that every convex function is quasiconvex, but not conversely, so this is a weaker condition. The following first-order characterization of the differentiable case is used below.

Lemma B.2 (First-order characterization)**.** *Let* f *be differentiable and quasiconvex on an open convex set* $\mathcal{C}$*. Then for all* $x, y \in \mathcal{C}$*,*

$$f(y) \le f(x) \implies \langle \nabla f(x), y - x \rangle \le 0. \tag{41}$$

This is the standard first-order characterization of differentiable quasiconvex functions (Boyd & Vandenberghe, 2004).

Quasiconvexity is also preserved under an invertible linear change of coordinates.

Lemma B.3 (Invariance under linear maps)**.** *Let M be invertible and $\widetilde{f} = f \circ M^{-1}$. If f is quasiconvex on an open convex set $\mathcal{C}$, then $\widetilde{f}$ is quasiconvex on the open convex set $M\mathcal{C}$.*

Proof. An invertible linear map sends open convex sets to open convex sets, so $M\mathcal{C}$ is open and convex.

For any α, the sublevel set $\{\widetilde{x} : \widetilde{f}(\widetilde{x}) \le \alpha\} = M\{x : f(x) \le \alpha\}$ is the image of a convex set under a linear map, hence convex. □

Quasiconvexity connects to the SAM perturbation through the following alignment property.

Lemma B.4 (Gradient alignment under quasiconvexity)**.** *Let f be differentiable and quasiconvex on an open convex set $\mathcal{C}$, with $x, x^+ \in \mathcal{C}$, $g = \nabla f(x) \neq 0$, $g^+ = \nabla f(x^+) \neq 0$, and $x^+ = x + \rho\, g/\left\|g\right\|_2$ for some $\rho > 0$. Then $\langle g, g^+\rangle > 0$.*

Proof. Step 1: $f(x^+) > f(x)$.

If instead $f(x^+) \le f(x)$, then Lemma B.2 gives $\langle g, x^+ - x\rangle \le 0$, which contradicts

$$\left\langle g, x^+ - x\right\rangle = \rho \left\|g\right\|_2 > 0. \tag{42}$$

Step 2: $\langle g, g^+\rangle > 0$.

Let $\mathcal{S} = \{z \in \mathcal{C} : f(z) \le f(x^+)\}$, which is convex since f is quasiconvex. By Lemma B.2, $\langle g^+, z - x^+\rangle \le 0$ for all $z \in \mathcal{S}$, so the hyperplane through x^+ with normal g^+ supports $\mathcal{S}$.

By Step 1, $f(x) < f(x^+)$, so x lies in the interior of $\mathcal{S}$ and $\langle g^+, x - x^+\rangle \le 0$. Since the supporting hyperplane cannot pass through an interior point of $\mathcal{S}$, the inequality is strict: $\langle g^+, x - x^+\rangle < 0$. As $x - x^+ = -\rho\, g/\left\|g\right\|_2$, we conclude $\langle g, g^+\rangle > 0$.

□

B.5 Proof of Theorem 5.9

Since $x_t^+ = x_t + \epsilon_t = x_t + \rho\, g_t/\left\|g_t\right\|_2$, applying Lemma B.4 to f_B with $\mathcal{C} = \mathcal{Q}$ and the points x_t, x_t^+ gives $\left\langle g_t, g_t^+\right\rangle > 0$.

It remains to establish descent. By L_B-smoothness of f_B,

$$f_B(x_{t+1}) \le f_B(x_t) - \eta_t \left\langle g_t, g_t^+\right\rangle + \frac{L_B \eta_t^2}{2}\left\|g_t^+\right\|_2^2 = f_B(x_t) - \eta_t\left(\left\langle g_t, g_t^+\right\rangle - \frac{L_B \eta_t}{2}\left\|g_t^+\right\|_2^2\right), \tag{43}$$

so $f_B(x_{t+1}) < f_B(x_t)$ whenever $0 < \eta_t < 2\left\langle g_t, g_t^+\right\rangle / (L_B \left\|g_t^+\right\|_2^2)$.

Finally, since $\mathcal{Q}$ is open and $x_t \in \mathcal{Q}$, and $x_{t+1} = x_t - \eta_t g_t^+ \to x_t$ as $\eta_t \to 0$, we have $x_{t+1} \in \mathcal{Q}$ for all sufficiently small η_t. Both conditions hold once η_t is small enough. □

B.6 Proof of Corollary 5.10

Alignment. Consider the scaled coordinates $\widetilde{x} = P_t^{1/2} x$ and $\widetilde{f}_B = f_B \circ P_t^{-1/2}$. Since $P_t^{1/2}$ is invertible, Lemma B.3 shows $\widetilde{f}_B$ is quasiconvex on the open convex set $\widetilde{\mathcal{Q}} = P_t^{1/2}\mathcal{Q}$, and $x_t, x_t^+ \in \mathcal{Q}$ gives $\widetilde{x}_t, \widetilde{x}_t^+ \in \widetilde{\mathcal{Q}}$. By Proposition 4.1, in these coordinates the PSAM perturbation is the standard SAM perturbation for $\widetilde{f}_B$ with radius $\widetilde{\rho}_t$, so Lemma B.4 applied to $\widetilde{f}_B$ yields

$$\left\langle \nabla \widetilde{f}_B(\widetilde{x}_t), \nabla \widetilde{f}_B(\widetilde{x}_t^+)\right\rangle = \left\langle P_t^{-1/2} g_t, P_t^{-1/2} g_t^+\right\rangle = g_t^\top P_t^{-1} g_t^+ > 0. \tag{44}$$

Descent. The PSAM update $x_{t+1} = x_t - \eta_t P_t^{-1} g_t^+$ is taken in the original coordinates, where f_B is L_B-smooth. Hence

$$f_B(x_{t+1}) \le f_B(x_t) - \eta_t \, g_t^\top P_t^{-1} g_t^+ + \frac{L_B \eta_t^2}{2} \left\| P_t^{-1} g_t^+ \right\|_2^2, \tag{45}$$

so $f_B(x_{t+1}) < f_B(x_t)$ whenever $0 < \eta_t < 2\, g_t^\top P_t^{-1} g_t^+ / (L_B \left\| P_t^{-1} g_t^+ \right\|_2^2)$, which is positive by the alignment above.

Containment. Since $\mathcal{Q}$ is open, $x_t \in \mathcal{Q}$, and $x_{t+1} \to x_t$ as $\eta_t \to 0$, we have $x_{t+1} \in \mathcal{Q}$ for all sufficiently small η_t. All three conditions hold once η_t is small enough. □

B.7 A QUADRATIC REALIZATION OF THE MISMATCH

We exhibit a function for which the preconditioned update with a standard SAM perturbation violates the descent condition. Let

$$f_B(x) = \tfrac{1}{2} x^\top A x, \qquad A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}. \tag{46}$$

The eigenvalues of A are 2 and 4, so f_B is strongly convex on $\mathbb{R}^2$, hence quasiconvex, and $L_B = 4$. Taking $\mathcal{Q} = \mathbb{R}^2$ satisfies the assumptions of Theorem 5.9.

At $x_t = (2, -1)^\top$ we have $g_t = (5, -1)^\top$. Choosing $\rho = \|g_t\|_2 = \sqrt{26}$ gives $x_t^+ = (7, -2)^\top$ and $g_t^+ = (19, 1)^\top$, so that with $P_t^{-1} = \mathrm{Diag}(0.1, 10)$,

$$\left\langle g_t, g_t^+ \right\rangle = 94 > 0, \qquad \left\langle g_t, P_t^{-1} g_t^+ \right\rangle = -0.5 < 0. \tag{47}$$

The alignment holds in the Euclidean geometry but fails in the preconditioned one.

C DIAGNOSTICS FOR THE PSAM ANALYSIS

The constants $\widetilde{L}$, G, and $d_{\max}$ of Section 5 are uniform bounds on quantities that vary along training. We record those quantities at each step, together with the effective radius $\widetilde{\rho}_t$ of equation 7, along the PSAM runs that supply the preconditioner drift of Section 6; see Appendix D.1 for the experimental setup. Assumptions 5.3 and 5.4 bound $\left\| P_t^{-1} g_t^+ \right\|_2$ and the largest eigenvalue $\lambda_{\max}(P_t)$, which we measure directly. Assumption 5.2 instead bounds a difference scaled by $\widetilde{\rho}_t$, so we report

$$\widetilde{L}_t := \frac{\left\| P_t^{-1/2} (g_t^+ - g_t) \right\|_2}{\widetilde{\rho}_t}, \tag{48}$$

the smallest value satisfying it at step t, so that the assumption holds with $\widetilde{L} = \sup_t \widetilde{L}_t$.

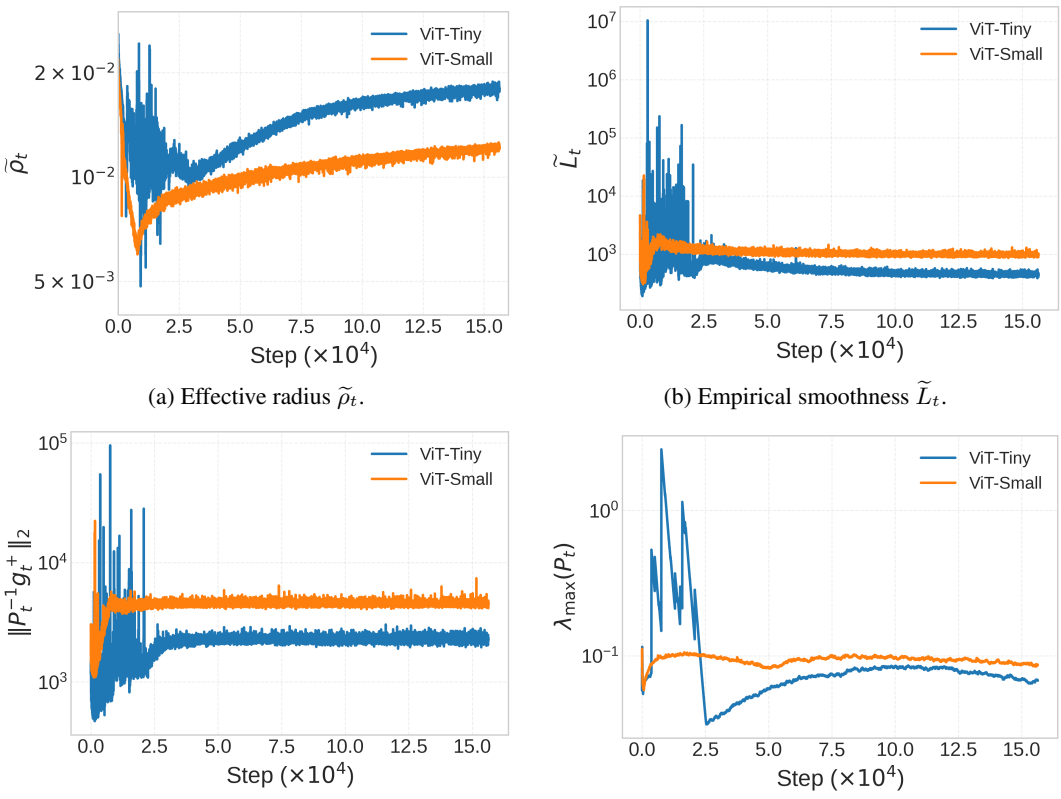
The results are shown in Figure 5. All four quantities show a similar pattern to the preconditioner drift in Section 6: pronounced fluctuations early in training give way to more stable behavior by the middle of training. The assumptions of Section 5 are therefore plausible once the preconditioner has settled, though not over the early phase.

D EXPERIMENTAL DETAILS

D.1 IMAGE CLASSIFICATION

For image classification, we evaluate CIFAR-100. We use two Vision Transformer sizes, ViT-Tiny (`vit_tiny_patch16_224`) and ViT-Small (`vit_small_patch16_224`) from the timm library, both trained from scratch with a 32×32 input and a 4×4 patch size.

We use AdamW as the base optimizer with betas $(0.9, 0.999)$, $c = 10^{-8}$, weight decay of 5×10^{-4}, and a constant learning rate of 10^{-4} for ViT-Tiny and 5×10^{-5} for ViT-Small. Training uses automatic mixed precision with BF16 and a batch size of 64. For augmentation, we apply random crop, random horizontal flip, and the AutoAugment policy, with CIFAR-100 normalization applied



(a) Effective radius $\widetilde{\rho}_t$.

(b) Empirical smoothness $\widetilde{L}_t$.

(c) Update norm $\left\|P_t^{-1} g_t^+\right\|_2$.

(d) Largest eigenvalue $\lambda_{\max}(P_t)$.

Figure 5: Quantities entering the analysis of Section 5, measured during a single PSAM training run ($\rho = 1$) for ViT-Tiny and ViT-Small on CIFAR-100. Panels (b), (c), and (d) correspond to Assumptions 5.2, 5.3, and 5.4, respectively. All curves use a 10-step rolling mean.

Table 3: CIFAR-100 final test accuracy (%) versus ρ for ViT-Tiny, over seeds $1, 2, 3$ (mean $\pm$ std). AdamW baseline: 62.47 ± 0.51. WS-SAM/WS-PSAM use a warm-start ratio $\tau = 0.4$.

ρ	SAM	PSAM	WS-SAM	WS-PSAM
0.05	63.73 ± 0.35	64.04 ± 0.42	64.43 ± 0.23	64.53 ± 0.36
0.1	63.76 ± 0.45	64.14 ± 0.72	65.38 ± 0.11	65.72 ± 0.23
0.2	63.75 ± 0.57	64.10 ± 0.31	65.99 ± 0.15	66.65 ± 0.19
0.5	64.48 ± 1.57	63.33 ± 0.08	67.45 ± 0.45	67.64 ± 0.52
1	52.80 ± 6.82	50.09 ± 2.10	68.99 ± 0.24	$\mathbf{69.87 \pm 0.18}$
2	45.13 ± 4.53	1.00 ± 0.00	59.62 ± 1.29	29.00 ± 4.81

at both training and test time. We train for 200 epochs and run three seeds, $1, 2, 3$, reporting the final test accuracy.

For SAM-type methods, we sweep $\rho \in \{0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$. WS-SAM and WS-PSAM use a warm start over the first 40% of training by default. For the warm-start analysis, we additionally sweep the warm-start ratio from 0.0 to 1.0 for WS-PSAM at $\rho = 1.0$.

The preconditioner-drift experiment (Figure 3) uses the same setup with $\rho = 1$ and seed 0.

Large-radius behavior of WS-PSAM in vision. For both ViT-Tiny and ViT-Small, WS-PSAM degrades sharply at $\rho = 2$, whereas WS-SAM remains comparatively stable (Tables 3 and 4). Figure 6 shows how this arises during training: at $\rho = 2$ both methods jump to a similar loss when the perturbation is switched on, but WS-SAM comes back down, though not to the level of the $\rho = 1$ runs, while WS-PSAM remains at high loss for the rest of training. Section 5.3 assumes $x_t \in \mathcal{Q}$

Table 4: CIFAR-100 final test accuracy (%) versus ρ for ViT-Small, over seeds $1, 2, 3$ (mean $\pm$ std). AdamW baseline: 63.22 ± 0.28. WS-SAM/WS-PSAM use a warm-start ratio $\tau = 0.4$.

ρ	SAM	PSAM	WS-SAM	WS-PSAM
0.05	64.98 ± 0.54	65.60 ± 0.72	65.78 ± 0.13	65.74 ± 0.09
0.1	66.14 ± 0.36	66.78 ± 0.45	66.65 ± 0.22	66.83 ± 0.36
0.2	66.74 ± 0.11	66.29 ± 0.25	67.42 ± 0.22	67.92 ± 0.44
0.5	66.37 ± 0.58	63.77 ± 1.46	67.68 ± 0.32	69.08 ± 0.35
1	62.85 ± 0.93	46.73 ± 2.69	68.36 ± 2.09	$\mathbf{71.42 \pm 0.28}$
2	55.98 ± 6.02	27.55 ± 5.88	67.17 ± 0.29	36.22 ± 7.79

Table 5: CIFAR-100 final test accuracy (%) of WS-PSAM at $\rho = 1$ as the warm-start ratio τ varies, over seeds $1, 2, 3$ (mean $\pm$ std). $\tau = 0$ corresponds to PSAM and $\tau = 1$ to AdamW.

τ	ViT-Tiny	ViT-Small
0.0	50.09 ± 2.10	46.73 ± 2.69
0.1	67.54 ± 0.23	69.75 ± 0.12
0.2	68.98 ± 0.47	70.87 ± 0.10
0.3	69.68 ± 0.32	71.26 ± 0.20
0.4	$\mathbf{69.87 \pm 0.18}$	$\mathbf{71.42 \pm 0.28}$
0.5	69.83 ± 0.20	71.14 ± 0.06
0.6	69.82 ± 0.14	71.30 ± 0.24
0.7	69.79 ± 0.26	69.58 ± 0.37
0.8	69.69 ± 0.34	68.29 ± 0.29
0.9	68.21 ± 0.21	68.10 ± 0.12
1.0	62.47 ± 0.51	63.22 ± 0.28

and notes that this is more plausible at low loss, so a run that stays at high loss may fall outside the admissible-radius argument. A likely cause is that the preconditioner is updated with g_t^+ rather than g_t, and the resulting disturbance grows with ρ. Both methods share this, but the PSAM perturbation is itself built from the preconditioner, so PSAM receives the same disturbance through two paths rather than one.

D.2 Machine translation

For machine translation, we evaluate IWSLT14 German-English using the standard setup and evaluation pipeline provided by fairseq. We use a Transformer architecture with a 6-layer encoder and a 6-layer decoder, an embedding dimension of 512, an FFN dimension of 1024, 4 attention heads, shared decoder input/output embeddings, and approximately 39.5M parameters.

We use Adam as the base optimizer with a learning rate of 5×10^{-4}, betas $(0.9, 0.98)$, weight decay of 10^{-4}, and a cosine learning-rate schedule with linear warmup over the first 5% of training. We use dropout of 0.3, label smoothing of 0.1, and a maximum of 4096 tokens per batch. Training uses automatic mixed precision with FP16. We train for 70 epochs and run three seeds, $1, 2, 3$.

For each ρ, we report the test BLEU obtained from the final checkpoint. For SAM-type methods, we sweep $\rho \in \{0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0\}$. WS-SAM and WS-PSAM use a warm start over the first 20% of training.

D.3 Language modeling

We evaluate two GPT-2 model sizes: GPT-2 (30M) on WikiText-2 and GPT-2 (124M) on WikiText-103. Both models use the GPT-2 tokenizer and are implemented using Hugging Face Transformers.

For both model sizes, we use AdamW as the base optimizer with betas $(0.9, 0.95)$, weight decay of 0.1, and a cosine learning-rate schedule with linear warmup over the first 5% of training. Training uses automatic mixed precision with FP16. WS-SAM and WS-PSAM use a warm start over the first 20% of training.

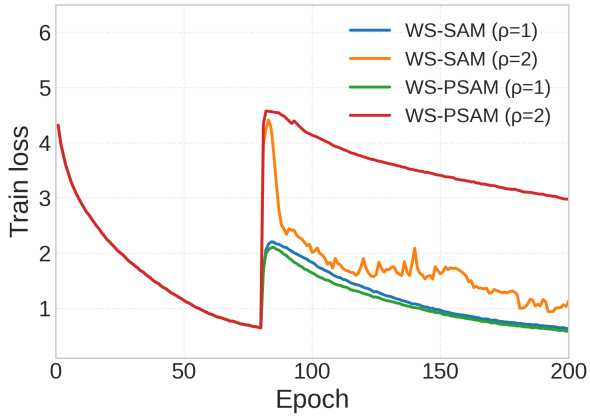


Figure 6: Training loss on CIFAR-100 with ViT-Tiny (seed 1). The perturbed gradient is switched on at epoch 80 ($\tau = 0.4$).

Table 6: IWSLT14 De-En final test BLEU versus ρ, over seeds $1, 2, 3$ (mean $\pm$ std). Adam baseline: 34.28 ± 0.26. WS-SAM/WS-PSAM use a warm-start ratio $\tau = 0.2$. Entries near 0 indicate training collapse.

ρ	SAM	PSAM	WS-SAM	WS-PSAM
0.1	34.26 ± 0.06	34.27 ± 0.16	34.22 ± 0.05	34.24 ± 0.10
0.2	34.33 ± 0.03	34.39 ± 0.03	34.33 ± 0.15	34.29 ± 0.22
0.5	2.00 ± 1.51	34.42 ± 0.08	34.48 ± 0.05	34.28 ± 0.12
1.0	0.00 ± 0.00	33.52 ± 0.87	34.66 ± 0.09	34.49 ± 0.10
2.0	0.00 ± 0.00	0.01 ± 0.01	23.45 ± 20.03	34.46 ± 0.13
5.0	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	34.59 ± 0.14
10.0	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	$\mathbf{34.79 \pm 0.04}$
20.0	0.01 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	34.69 ± 0.16

GPT-2 (30M). The model consists of 6 layers, 6 attention heads, an embedding dimension of 384, and a sequence length of 128. We train for 2,500 steps with a batch size of 64, processing approximately 20.48M tokens in total. We use a learning rate of 10^{-3}, dropout of 0.1, and gradient clipping of 1.0. We run five seeds, $1, 2, 3, 4, 5$, and report the final validation perplexity.

For SAM-type methods, we sweep $\rho \in \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0\}$.

GPT-2 (124M). The model consists of 12 layers, 12 attention heads, an embedding dimension of 768, and a sequence length of 1024. We train for 10,000 steps with a batch size of 64, processing approximately 655.36M tokens in total. We use a learning rate of 2×10^{-3} and do not use dropout or gradient clipping. We run three seeds, $1, 2, 3$, and report the final validation perplexity.

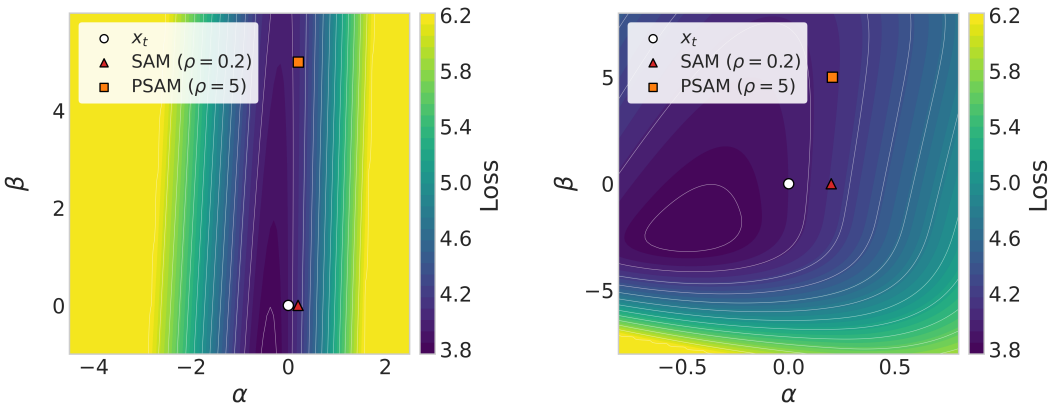
For all SAM-type methods, we use the same ρ grid as in the GPT-2 (30M) experiments.

Table 7: GPT-2 (30M) on WikiText-2 final validation perplexity versus ρ, over seeds 1–5 (mean $\pm$ std). AdamW baseline: 160.73 ± 0.80. WS-SAM/WS-PSAM use a warm-start ratio $\tau = 0.2$.

ρ	SAM	PSAM	WS-SAM	WS-PSAM
0.005	160.42 ± 0.59	159.93 ± 0.83	160.34 ± 0.86	160.46 ± 0.90
0.01	160.23 ± 0.68	159.94 ± 0.76	160.20 ± 0.84	160.44 ± 0.90
0.02	160.29 ± 0.79	160.31 ± 0.63	159.92 ± 0.81	160.39 ± 0.90
0.05	161.16 ± 0.54	160.05 ± 0.75	159.23 ± 0.78	160.23 ± 0.89
0.1	163.54 ± 0.71	159.89 ± 0.72	158.54 ± 0.65	159.98 ± 0.89
0.2	168.42 ± 0.68	159.51 ± 0.85	158.37 ± 0.69	159.52 ± 0.91
0.5	232.41 ± 15.50	158.60 ± 0.71	161.86 ± 0.54	158.44 ± 0.91
1.0	357.87 ± 15.24	157.30 ± 0.31	173.54 ± 0.92	157.24 ± 0.88
2.0	483.62 ± 9.36	157.90 ± 0.68	270.99 ± 16.17	155.92 ± 0.84
5.0	811.47 ± 159.25	204.14 ± 14.30	706.68 ± 273.04	$\mathbf{155.61 \pm 0.74}$
10.0	1354.71 ± 102.97	266.48 ± 7.09	2446.77 ± 1179.78	161.29 ± 0.64

Table 8: GPT-2 (124M) on WikiText-103 final validation perplexity versus ρ, over seeds $1, 2, 3$ (mean $\pm$ std). AdamW baseline: 18.28 ± 0.05. WS-SAM/WS-PSAM use a warm-start ratio $\tau = 0.2$.

ρ	SAM	PSAM	WS-SAM	WS-PSAM
0.005	18.29 ± 0.04	18.27 ± 0.03	18.29 ± 0.07	18.29 ± 0.08
0.01	18.23 ± 0.02	18.27 ± 0.02	18.28 ± 0.06	18.28 ± 0.03
0.02	18.24 ± 0.05	18.24 ± 0.02	18.29 ± 0.02	18.28 ± 0.02
0.05	18.30 ± 0.03	18.26 ± 0.10	18.29 ± 0.06	18.32 ± 0.04
0.1	18.69 ± 0.27	18.26 ± 0.04	18.29 ± 0.07	18.29 ± 0.12
0.2	19.48 ± 0.17	18.31 ± 0.01	18.24 ± 0.09	18.30 ± 0.09
0.5	378.39 ± 20.91	18.28 ± 0.05	18.23 ± 0.10	18.24 ± 0.03
1.0	448.20 ± 34.42	18.47 ± 0.44	18.20 ± 0.07	18.22 ± 0.03
2.0	453.59 ± 98.91	18.98 ± 0.22	18.66 ± 0.04	18.18 ± 0.09
5.0	332.07 ± 52.56	19.48 ± 0.27	43.14 ± 27.39	$\mathbf{18.09 \pm 0.03}$
10.0	330.61 ± 30.95	21.08 ± 0.32	214.84 ± 7.27	18.36 ± 0.09



(a) Common scale.

(b) β range $10\times$ wider.

Figure 7: Loss on the plane spanned by the SAM and PSAM perturbation directions at the end of a WS-PSAM run on GPT-2 (30M), seed 1. Both panels show the same plane, (a) with the two axes on a common scale and (b) over a wider range along e_2, drawn with the β axis compressed by a factor of 10 relative to α. Markers: x_t (circle), SAM at $\rho = 0.2$ (triangle), PSAM at $\rho = 5$ (square).

D.4 Loss plane along the perturbation directions

We visualize the loss landscape of a trained model on the plane spanned by the SAM and PSAM perturbation directions.

We take x_t to be the final parameters of a WS-PSAM run ($\rho = 5$) on GPT-2 (30M) with seed 1, and let g_t be the gradient of a training mini-batch at that point. With $e_1 = g_t/\|g_t\|_2$ and e_2 obtained from $P_t^{-1} g_t$ by removing its e_1 component and normalizing, we evaluate the loss on

$$x(\alpha, \beta) = x_t + \alpha e_1 + \beta e_2. \tag{49}$$

Every grid point is evaluated on the mini-batch that defines the plane. We mark the two perturbations at the radii that perform best on this model, $\rho = 0.2$ for WS-SAM and $\rho = 5$ for WS-PSAM (Table 2).

In Figure 7(a), the two perturbation directions are nearly orthogonal ($87.6°$), matching the cosine similarity of 0.040 in Table 2, so the PSAM perturbation lies almost along e_2. Since the shape of the level sets is hard to see in (a), we compress the vertical axis $10\times$ in (b), where they are indeed seen to be convex-like. The SAM and PSAM perturbations raise the loss from 3.93 to 4.10 and 4.07, respectively, yet their radii differ by a factor of 25. This is similar to the situation in Figure 2: PSAM may need a larger radius than SAM to leave a convex sublevel set.