Journal of Machine Learning Research 24 (2023) 1-36 Submitted 4/23; Revised 10/23; Published 10/23

The Dynamics of Sharpness-Aware Minimization: Bouncing Across Ravines and Drifting Towards Wide Minima

Peter L. Bartlett[*] PETERBARTLETT@GOOGLE.COM
Philip M. Long PLONG@GOOGLE.COM
Olivier Bousquet OBOUSQUET@GOOGLE.COM
Google
1600 Amphitheatre Parkway
Mountain View, CA 94040

Editor: Ohad Shamir

Abstract

We consider Sharpness-Aware Minimization (SAM), a gradient-based optimization method for deep networks that has exhibited performance improvements on image and language prediction problems. We show that when SAM is applied with a convex quadratic objective, for most random initializations it converges to a cycle that oscillates between either side of the minimum in the direction with the largest curvature, and we provide bounds on the rate of convergence.

In the non-quadratic case, we show that such oscillations effectively perform gradient descent, with a smaller step-size, on the spectral norm of the Hessian. In such cases, SAM's update may be regarded as a third derivative—the derivative of the Hessian in the leading eigenvector direction—that encourages drift toward wider minima.

Keywords: Non-convex optimization, wide minima, sharpness-aware minimization.

1. Introduction

The broad practical impact of deep learning has heightened interest in many of its surprising characteristics: simple gradient methods applied to deep neural networks seem to efficiently optimize nonconvex criteria, reliably giving a near-perfect fit to training data, but exhibiting good predictive accuracy nonetheless (see Bartlett et al., 2021). Optimization methodology is widely believed to affect statistical performance by imposing some kind of implicit regularization, and there has been considerable effort devoted to understanding the behavior of optimization methods and the nature of solutions that they find. For instance, Barrett and Dherin (2020) and Smith et al. (2020) show that discrete-time gradient descent and stochastic gradient descent can be viewed as gradient flow methods applied to penalized losses that encourage smoothness, and Soudry et al. (2018) and Azulay et al. (2021) identify the implicit regularization imposed by gradient flow in specific examples, including linear networks.

[*] Also affiliated with University of California, Berkeley.

We consider *Sharpness-Aware Minimization* (SAM), a recently introduced (Foret et al., 2020) gradient optimization method that has exhibited substantial improvements in prediction performance for deep networks applied to image classification (Foret et al., 2020) and NLP (Bahri et al., 2022) problems.

In introducing SAM, Foret *et al* motivate it using a minimax optimization problem

$$\min_{w} \max_{\|\epsilon\| \leq \rho} \ell(w + \epsilon), \tag{1}$$

where $\ell : \mathbb{R}^d \to \mathbb{R}$ is an empirical loss defined on the parameter space $\mathbb{R}^d$, $\|\cdot\|$ is the Euclidean norm on the parameter space, and ρ is a scale parameter. By viewing the difference

$$\max_{\|\epsilon\| \leq \rho} \ell(w + \epsilon) - \ell(w)$$

as a measure of the sharpness of the empirical loss ℓ at the parameter value w, the criterion in (1) allows a trade-off between the empirical loss and the sharpness,

$$\max_{\|\epsilon\| \leq \rho} \ell(w + \epsilon) = \ell(w) + \underbrace{\max_{\|\epsilon\| \leq \rho} \ell(w + \epsilon) - \ell(w)}_{\text{sharpness}}.$$

In practice, SAM works with a simplification based on gradient measurements, starting with an initial parameter vector $w_0 \in \mathbb{R}^d$ and updating the parameters at iteration t via

$$w_{t+1} = w_t - \eta \nabla \ell \left(w_t + \rho \frac{\nabla \ell(w_t)}{\|\nabla \ell(w_t)\|} \right), \tag{2}$$

where η is a step-size parameter. Our goal in this paper is to understand the nature of the solutions that the SAM updates (2) lead to.

In Sections 3 and 4, we consider SAM with a convex quadratic criterion. The key insight is that it is equivalent to a gradient descent method for a certain non-convex criterion whose stationary points correspond to oscillations around the minimum in the directions of the eigenvectors of the Hessian of the loss. The only stable stationary point corresponds to the leading eigenvector direction: ‘bouncing across the ravine’. (Notice that this is not the solution to the motivating minimax optimization problem (1), which is the minimum of the quadratic criterion.)

In Section 5, we analyze one of SAM’s updates near a smooth minimum of the loss function ℓ with a positive semidefinite Hessian. For parameters corresponding to the solutions for the quadratic case, we see that the SAM updates can be decomposed into two components. There is a large component in the direction of the oscillation (bouncing across the ravine), and there is a smaller component in the orthogonal subspace that corresponds to descending the gradient of the spectral norm of the Hessian. Thus, SAM is able to drift towards wide minima by exploiting a specific third derivative (the gradient of the second derivative in the leading eigenvalue direction) with only two gradient computations per iteration. In Section 7, we present some open problems, the most important of which is elucidating the relationship between wide minima of empirical loss and statistical performance.

2. Additional Related Work

Du et al. (2022) proposed a more computationally efficient variant of SAM. Beugnot et al. (2022) studied the effect of a large learning rate with early stopping on spectrum of the Hessian in the case of quadratic loss.

Cohen et al. (2020) provided a variety of natural settings where, empirically, when neural networks are trained with batch gradient descent and a fixed learning rate η, the spectral norm of the Hessian tends toward $2/\eta$, the "edge of stability". Here, if the gradient is aligned with the principal direction of the Hessian, the solution "bounces across the ravine", as in the analysis of this paper. A number of theoretical treatments of this phenomenon have since been proposed (Ahn et al., 2022; Arora et al., 2022; Damian et al., 2022). The most closely related of those to this paper is the work of Damian et al. (2022), who also described conditions under which "bouncing across the ravine" tends to decrease the spectral norm of the Hessian.

In independent work posted to arXiv after the initial version of this paper, Wen et al. (2023) performed a variety of analyses of SAM and some related algorithms. Their results included showing that SAM almost surely converges in the limit in the convex quadratic case for suitably small values of various problem parameters, along with asymptotic analysis showing that, under assumptions on the loss function and on the existence of a suitable manifold of loss minimizers, once SAM gets close enough to this manifold, it approximately tracks the path on the manifold of gradient flow with respect to the spectral norm of the Hessian. Our analysis of the convex quadratic case reveals an equivalence to gradient descent on a non-convex objective and uses it, under explicit conditions on the various problem parameters, to give explicit convergence rates to a limiting set. And under an explicit smoothness condition on a non-quadratic loss, we show that the SAM update from this set corresponds to gradient descent in the spectral norm of the Hessian. Wen et al. (2023) also extended their asymptotic analysis to a stochastic version of SAM, in which both gradients at each step are estimated from a single training example, showing that the approximate gradient flow in this case is with respect to the trace of the Hessian.

3. SAM with Quadratic Loss: Bouncing Across Ravines

We first consider the application of SAM to minimize a convex quadratic objective ℓ. Without loss of generality, we assume that the minimum of ℓ is at zero, the eigenvectors of ℓ's Hessian are the coordinate axes, and the eigenvalues are sorted by the indices of the eigenvectors. Accordingly, for $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 \geq \cdots \geq \lambda_d > 0$, we consider loss $\ell(w) = \frac{1}{2} w^\top \Lambda w$. Then $\nabla \ell(w) = \Lambda w$ and SAM sets

$$\begin{aligned} w_{t+1} &= w_t - \eta \nabla \ell \left(w_t + \rho \frac{\nabla \ell(w_t)}{\|\nabla \ell(w_t)\|} \right) \\ &= \left(I - \eta \Lambda - \frac{\eta \rho}{\|\Lambda w_t\|} \Lambda^2 \right) w_t. \end{aligned} \tag{3}$$

The following is our main result.

Theorem 1 *There are polynomials p and p' and an absolute constant c such that the following holds. For any eigenvalues $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_d > 0$, loss $\ell(w) = \frac{1}{2} w^\top \Lambda w$ with*

$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, any neighborhood size $\rho > 0$, any step size $0 < \eta < \frac{1}{2\lambda_1}$, and any $\delta > 0$, if w_0 is sampled from a continuous probability distribution over $\mathbb{R}^d$

- *whose density is bounded above by $A \in \mathbb{R}$, and*
- *for $R > \eta\rho\lambda_1$ and $q > 0$, with probability at least $1 - \delta$, $\|w_0\| \leq R$ and $w_{0,1}^2 \geq q$,*

and $w_1, w_2, \ldots$ are obtained through the SAM update (2), then, if $\kappa = \lambda_1/\lambda_d$, for all

$$\epsilon < p'(1/\lambda_1, \lambda_d, \eta, \rho, \delta, 1/\rho, 1/A, 1/R, q),$$

with probability $1 - 2\delta$, for all

$$t \geq \left(\frac{\kappa^5}{\eta\lambda_d \min\left\{\eta\lambda_d, \lambda_1^2/\lambda_2^2 - 1\right\}} + d \right) p\left(\log\left(\frac{1}{\epsilon}\right)\right)$$

one of the following holds:

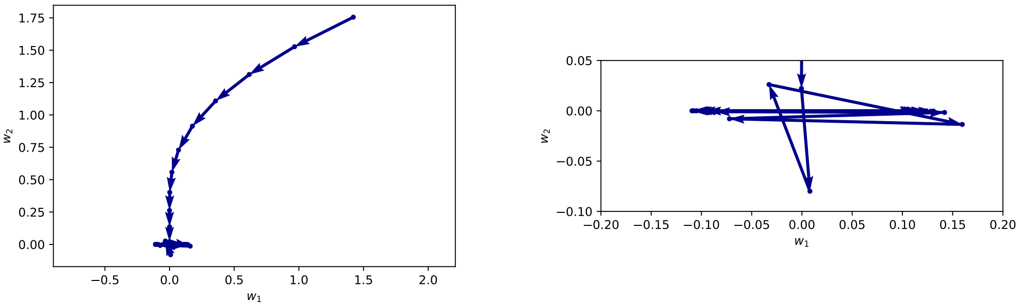
- *$\|w_t - \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \leq \epsilon$ and $\|w_{t+1} + \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \leq \epsilon$, or*
- *$\|w_t + \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \leq \epsilon$ and $\|w_{t+1} - \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \leq \epsilon$.*

Theorem 1 has the following corollary.

Theorem 2 *For any eigenvalues $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_d > 0$, any neighborhood size $\rho > 0$, and any step size $0 < \eta < \frac{1}{2\lambda_1}$, if w_0 is sampled from a continuous probability distribution over $\mathbb{R}^d$ with $\mathbf{E}[\|w_0\|^2] < \infty$, then, almost surely, for all $\epsilon > 0$, for all large enough t, the iterates of SAM applied to the quadratic loss $\ell(w) = \frac{1}{2} w^\top \mathrm{diag}(\lambda_1, \ldots, \lambda_d) w$ satisfy:*

- *$\|w_t - \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \leq \epsilon$ and $\|w_{t+1} + \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \leq \epsilon$, or*
- *$\|w_t + \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \leq \epsilon$ and $\|w_{t+1} - \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \leq \epsilon$.*

Our analysis shows that, when SAM is initialized far from the optimum, training proceeds in two stages. Early, the objective function is reduced exponentially fast, with the most rapid progress made in the directions with highest variance. This can be seen, for example, in Figure 1a, which plots the first 30 iterates of SAM initialized at $(2, 2)$ in the case that $\lambda_1 = 1$ and $\lambda_2 = 1/2$, $\eta = 1/5$ and $\rho = 1$. After a certain point, however, SAM's iterates "overshoot" in the direction of highest variance, as can be seen in Figure 1b, which is the same as Figure 1a, except zoomed in to the region near the origin, where the details of the later iterates can be seen. During this second phase, the share of the length of the parameter vector in the first component increases, and the process converges to the oscillation described in Theorem 2. Note that, as illustrated in Figure 1a, due to the normalization by $\|w_t\|$, the parameter vector can jump away from a position very close to the origin, with a correspondingly very small loss. However, as we will see, the training process makes steady progress with respect to a potential function that we will define in Section 4.3.



(a) All the iterates

(b) The iterates close to the origin

Figure 1: The first 30 iterates of SAM, initialized at $(2, 2)$ with $\lambda_1 = 1$ and $\lambda_2 = 1/2$, $\eta = 1/5$ and $\rho = 1$.

4. Proof of Theorem 1

In this section, we prove the following theorem, which implies Theorem 1. We denote $\max\{z, 0\}$ by $[z]_+$.

Theorem 3 *There is an absolute constant c such that, for any eigenvalues $\lambda_1 > \lambda_2 \geq ... \geq \lambda_d > 0$, loss $\ell(w) = \frac{1}{2} w^\top \Lambda w$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, any neighborhood size $\rho > 0$, any initialization parameters $R, A, q > 0$, and any step size $0 < \eta < \frac{1}{2\lambda_1}$, for all $0 < \epsilon < \min\left\{\sqrt{\eta\lambda_1/2}, 1/(2\rho\lambda_1), \eta\rho\lambda_1^2/2\right\}$, for all $\delta > 0$, if w_0 is sampled from a continuous probability distribution over $\mathbb{R}^d$*

- *whose density is bounded above by $A \in \mathbb{R}$, and*
- *with probability at least $1 - \delta$, $\|w_0\| \leq R$ and $w_{0,1}^2 \geq q$,*

and $w_1, w_2, ...$ are obtained through the SAM update (2), then, with probability $1 - 2\delta$, for all

$$
\begin{aligned}
t \geq\ & \frac{6\lambda_1^5}{\eta\lambda_d^6 \min\left\{\eta\lambda_d, \frac{\lambda_1^2}{\lambda_2^2} - 1\right\}} \log\left(\frac{4}{\eta\lambda_1}\right) \\
&+ \frac{1}{\min\left\{\eta\lambda_d, \frac{\lambda_1^2}{\lambda_2^2} - 1\right\}} \left(\log\left(\frac{4(1 + \eta\rho\lambda_1^2)^2}{\lambda_d^2\epsilon^2}\right) + \log\left(\frac{R^2}{q}\right)\right) \\
&+ \frac{2\left[\log\left(\frac{R}{\eta\rho\lambda_1}\right)\right]_+}{\eta\lambda_d \min\left\{\eta\lambda_d, \frac{\lambda_1^2}{\lambda_2^2} - 1\right\}} \left(\log(2\lambda_1 R) + \frac{\left[\log\left(\frac{R}{\eta\rho\lambda_1}\right)\right]_+ \log\left(\frac{9 \cdot 6^{d+3} R^3}{(\eta\lambda_d)^{d+3}(\eta\rho\lambda_1)^3}\right)}{\eta\lambda_d}\right. \\
&\qquad\qquad \left. + \log\left(\frac{4\pi^{d/2}(4\eta\rho\lambda_1^2)^{d-1}\left[\log\left(\frac{R}{\eta\rho\lambda_1}\right)\right]_+ A}{\Gamma(d/2)\delta\eta\lambda_d}\right)\right) \\
&+ \frac{6}{\eta\lambda_1} \ln\left(\frac{2(1 + \eta\rho\lambda_1^2)}{\lambda_d\epsilon}\right)
\end{aligned}
$$

one of the following holds:

- $\|w_t - \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \le \epsilon$ *and* $\|w_{t+1} + \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \le \epsilon$*, or*
- $\|w_t + \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \le \epsilon$ *and* $\|w_{t+1} - \frac{\eta\rho\lambda_1 e_1}{2-\eta\lambda_1}\| \le \epsilon$*.*

The proof of Theorem 3 requires some lemmas, which we prove first. Throughout this section, we assume that $\eta\lambda_1 < 1/2$ and we highlight where the assumption $\lambda_1 > \lambda_2$ is used.

The evolution of the gradient $\nabla\ell(w_t) = \Lambda w_t$ plays a key role in the dynamics of SAM. To simplify expressions, we refer to it using the shorthand v_t. Substituting into the SAM update (3) for the quadratic loss gives

$$v_{t+1} = \left(I - \eta\Lambda - \frac{\eta\rho}{\|v_t\|}\Lambda^2\right) v_t,$$

so, for all $i \in [d]$ and all t, we have

$$\begin{aligned} v_{t+1,i} &= \left(1 - \eta\lambda_i - \frac{\eta\rho\lambda_i^2}{\|v_t\|}\right) v_{t,i} \\ &= (1-\eta\lambda_i)\left(\|v_t\| - \frac{\eta\rho\lambda_i^2}{1-\eta\lambda_i}\right)\frac{v_{t,i}}{\|v_t\|} \\ &= (1-\eta\lambda_i)\left(\|v_t\| - \gamma_i\right)\frac{v_{t,i}}{\|v_t\|}, \end{aligned}$$

where $\gamma_i := \eta\rho\lambda_i^2/(1-\eta\lambda_i)$.

We need the following technical lemma.

Lemma 4 *For $x > 0$, $0 \le a < b$, $\alpha > \beta \ge 0$, and $a\alpha \ge b\beta$, we have $a^2(x-\alpha)^2 > b^2(x-\beta)^2$ iff $x < (a\alpha + b\beta)/(a+b)$.*

Proof Substituting shows that $b^2(x-\beta)^2 - a^2(x-\alpha)^2 = 0$ at $x = (a\alpha + b\beta)/(a+b) \ge 0$. Also, $b^2(x-\beta)^2 - a^2(x-\alpha)^2 \le 0$ at $x = 0$, which shows that the other zero of this convex quadratic occurs at $x \le 0$. ■

4.1 Some properties

The following lemma identifies some properties of SAM with the convex quadratic criterion. It shows that the magnitudes of the components of the gradient vector v_t have fixed points under SAM's update when the gradients are in the eigenvector directions and at distance β_i from the minimum, it shows that the norm of v_t determines how the magnitudes of its components grow, both in absolute terms (where the critical values are the β_i) and relative to the first component (where the critical values are the α_i), and it shows that, for $b = \eta\rho\lambda_1^2$, the set $\{v : \|v\| \le b\}$ is absorbing. Recall that we have assumed that $\eta\lambda_1 < 1/2$.

Lemma 5 *For $i = 1, \ldots, d$, define*

$$\begin{aligned}
\gamma_i &= \frac{\eta\rho\lambda_i^2}{1-\eta\lambda_i}, \\
\beta_i &= \frac{1-\eta\lambda_i}{2-\eta\lambda_i}\gamma_i = \frac{\eta\rho\lambda_i^2}{2-\eta\lambda_i}, \\
\alpha_i &= \frac{(1-\eta\lambda_1)\gamma_1 + (1-\eta\lambda_i)\gamma_i}{1-\eta\lambda_1 + 1 - \eta\lambda_i} \\
b &= (1-\eta\lambda_1)\gamma_1 = \eta\rho\lambda_1^2.
\end{aligned}$$

We have

1. $\left\{(s_1, \ldots, s_d) : \exists v_t,\ \forall 1 \le i \le d,\ v_{t+1,i}^2 = v_{t,i}^2 = s_i\right\} = \{0\} \cup \bigcup_{i=1}^d \mathrm{co}\{\beta_i^2 e_j : \beta_j = \beta_i\}$, *where* $\mathrm{co}(S)$ *denotes the convex hull of a set* S *and* e_j *is the* j*th basis vector in* $\mathbb{R}^d$.
2. *For* $1 \le i \le d$, $v_{t+1,i}^2 < v_{t,i}^2$ *iff* $\|v_t\| > \beta_i$.
3. *Suppose* $\lambda_1 > \lambda_2$, $v_{t,1}^2 > 0$, *and* $v_{t,i}^2 > 0$. *Then for* $i \in \{2, \ldots, d\}$, $\frac{v_{t+1,1}^2}{v_{t+1,i}^2} > \frac{v_{t,1}^2}{v_{t,i}^2}$ *iff* $\|v_t\| < \alpha_i$.
4. $\beta_d \le \cdots \le \beta_1 \le \alpha_d \le \cdots \alpha_2 \le \alpha_1 = \gamma_1$ *and* $\beta_1 \le b$. *Furthermore, if* $\lambda_d > 0$ *then* $\beta_1 < \alpha_d$.
5. $\|v_t\| \le b$ *implies* $\|v_{t+1}\| \le b$.

Proof We have

$$v_{t+1,i}^2 = (1-\eta\lambda_i)^2 \left(\|v_t\| - \gamma_i\right)^2 \frac{v_{t,i}^2}{\|v_t\|^2},$$

and so, for all i, if $v_{t+1,i}^2 = v_{t,i}^2$, then, either $v_{t,i}^2 = 0$ or $(\|v_t\| - \gamma_i)^2 = \|v_t\|^2/(1-\eta\lambda_i)^2$. This quadratic equation has only one non-negative solution, $\|v_t\| = \beta_i$ (and for $\|v_t\| > \beta_i$ $v_{t+1,i}^2 < v_{t,i}^2$, proving part 2). And so if $v_{t,i} \neq 0$ for some i, then every $v_{t,j}$ with $\beta_j \neq \beta_i$ must have $v_{t,j} = 0$, and in that case, $\|v_t\|^2 = \sum_{j:\beta_j=\beta_i} v_{t,j}^2$. This proves part 1.

To see why part 5 is true, notice that, if $\|v_t\| \le b$, we have

$$\begin{aligned}
\|v_{t+1}\|^2 &= \sum_i (1-\eta\lambda_i)^2(\|v_t\| - \gamma_i)^2 \frac{v_{t,i}^2}{\|v_t\|^2} \\
&\le \sum_i (1-\eta\lambda_i)^2 \max\{\|v_t\|^2, \gamma_i^2\} \frac{v_{t,i}^2}{\|v_t\|^2} \\
&\le \max_i \left\{(1-\eta\lambda_i)^2 \max\{\|v_t\|^2, \gamma_i^2\}\right\} \\
&\le \max_i \left\{\max\{(1-\eta\lambda_i)^2 b^2, (1-\eta\lambda_i)^2\gamma_i^2\}\right\} \\
&= \max_i \left\{\max\{(1-\eta\lambda_i)^2(1-\eta\lambda_1)^2\gamma_1^2, (1-\eta\lambda_i)^2\gamma_i^2\}\right\} \\
&= \max_i \left\{(1-\eta\lambda_i)^2\gamma_i^2\right\} \\
&= (1-\eta\lambda_1)^2\gamma_1^2 = b^2.
\end{aligned}$$

For part 3, $v_{t+1,1}^2/v_{t,1}^2 > v_{t+1,i}^2/v_{t,i}^2$ iff

$$(1-\eta\lambda_1)^2(\|v_t\|-\gamma_1)^2 > (1-\eta\lambda_i)^2(\|v_t\|-\gamma_i)^2.$$

But because $1-\eta\lambda_1 < 1-\eta\lambda_2 \le 1-\eta\lambda_i$, $\gamma_1 > \gamma_2 \ge \gamma_i$, and $(1-\eta\lambda_1)\gamma_1 > (1-\eta\lambda_2)\gamma_2 \ge (1-\eta\lambda_i)\gamma_i$, we can apply Lemma 4, which shows that for $\|v_t\| > 0$, this is equivalent to $\|v_t\| < \alpha_i$.

To see part 4, first notice that, for $f(x) = x^2/(2-\eta x)$, $f'(x) \ge 0$ for all $x \in [0, 1/\eta)$, which implies that the β_i are non-increasing in i. Also,

$$\alpha_d = \frac{(1-\eta\lambda_1)\gamma_1 + (1-\eta\lambda_d)\gamma_d}{1-\eta\lambda_1+1-\eta\lambda_d} \ge \frac{1-\eta\lambda_1}{1-\eta\lambda_1+1-\eta\lambda_d}\gamma_1 \ge \frac{1-\eta\lambda_1}{2-\eta\lambda_1}\gamma_1 = \beta_1,$$

and the last inequality is strict iff $\lambda_d > 0$. Also, for $i < j$, $\gamma_i \ge \gamma_j$ and $1-\eta\lambda_i \le 1-\eta\lambda_j$, hence

$$\begin{aligned}\alpha_i &= \frac{(1-\eta\lambda_1)\gamma_1 + (1-\eta\lambda_i)\gamma_i}{1-\eta\lambda_1+1-\eta\lambda_i} \\ &\ge \frac{(1-\eta\lambda_1)\gamma_1 + (1-\eta\lambda_i)\gamma_j}{1-\eta\lambda_1+1-\eta\lambda_i} \\ &\ge \frac{(1-\eta\lambda_1)\gamma_1 + (1-\eta\lambda_j)\gamma_j}{1-\eta\lambda_1+1-\eta\lambda_j} \\ &= \alpha_j.\end{aligned}$$

Finally, $\beta_1 < (1-\eta\lambda_1)\gamma_1$ because $\eta\lambda_1 < 1$.

■

4.2 Early descent

In this section, we show that SAM rapidly descends towards the gradient ball $\|v_t\| \le b$ and that under our conditions on the initialization, it reaches this ball with the magnitude of the first component, $|v_{t,1}|$, bounded away from zero. We shall see in Section 4.4 that this ensures SAM, applied to the quadratic loss ℓ, converges to the leading eigenvector direction.

The following lemma shows that when the solution is far from the optimum, SAM rapidly descends toward the optimum and the relative magnitude of the first component of the gradient does not get too small.

Lemma 6 *Suppose that, for $R > 0$, we have $\|v_0\| \le R$. For any $t \ge T := [\log(R/b)]_+/(\eta\lambda_d)$, we have $\|v_t\| \le b$.*

Furthermore, if, for $\Delta > 0$, $|\|v_t\| - \gamma_1| \ge \Delta$ for all t, then there is a $T_0 \le T$ satisfying $\|v_{T_0}\| \le b$ and

$$\frac{\sum_{i=2}^d v_{T_0,i}^2}{v_{T_0,1}^2} \le \left(\frac{2R}{\Delta}\right)^{2T} \frac{\sum_{i=2}^d v_{0,i}^2}{v_{0,1}^2}. \tag{4}$$

Thus,

$$\log\left(\frac{\sum_{i=2}^d v_{T_0,i}^2}{v_{T_0,1}^2}\right) \le \frac{2}{\eta\lambda_d}\left[\log\left(\frac{R}{b}\right)\right]_+ \log\left(\frac{2R}{\Delta}\right) + \log\left(\frac{R^2}{v_{0,1}^2}\right).$$

Proof If $R \leq b$, the lemma is an obvious consequence of Part 5 of Lemma 5 ; assume for the rest of the proof that $R > b$.

Notice that $\|v_t\|^2 \geq (\|v_t\| - \gamma_i)^2$ if and only if $2\|v_t\| \geq \gamma_i$. But

$$b = (1 - \eta\lambda_1)\gamma_1 \geq \gamma_1/2 \geq \gamma_i/2.$$

Thus, for any $\|v_t\| \geq b$, we have

$$\begin{aligned}\|v_{t+1}\|^2 &= \sum_{i=1}^{d} (1 - \eta\lambda_i)^2 \left(\|v_t\| - \gamma_i\right)^2 \frac{v_{t,i}^2}{\|v_t\|^2} \\ &\leq \max_i (1 - \eta\lambda_i)^2 \|v_t\|^2 \\ &= (1 - \eta\lambda_d)^2 \|v_t\|^2. \end{aligned} \tag{5}$$

From Lemma 5, part 5, if $\|v_t\| \leq b$ then $\|v_{t'}\| \leq b$ for $t' \geq t$. Thus, for all t satisfying $(1 - \eta\lambda_d)^t \|v_0\| \leq b$, we have $\|v_t\| \leq b$. This is equivalent to

$$t \geq \frac{\log(\|v_0\|/b)}{\log(1/(1 - \eta\lambda_d))}.$$

Since $\log(1 - \eta\lambda_d) \leq -\eta\lambda_d$, it suffices if

$$t \geq T = \frac{[\log(R/b)]_+}{\eta\lambda_d}.$$

For the second part of the lemma, as long as $\|v_t\| \geq b$ we have

$$\begin{aligned}\frac{v_{t+1,i}^2}{v_{t+1,1}^2} &= \frac{(1 - \eta\lambda_i)^2 (\|v_t\| - \gamma_i)^2 v_{t,i}^2}{(1 - \eta\lambda_1)^2 (\|v_t\| - \gamma_1)^2 v_{t,1}^2} \\ &\leq \frac{(1 - \eta\lambda_i)^2 \|v_t\|^2 v_{t,i}^2}{(1 - \eta\lambda_1)^2 \Delta^2 v_{t,1}^2} \\ &\leq \frac{(1 - \eta\lambda_d)^2 R^2 v_{t,i}^2}{(1 - \eta\lambda_1)^2 \Delta^2 v_{t,1}^2}.\end{aligned}$$

Thus, if T_0 is the first iterate for which $\|v_{T_0}\| < b$, we have

$$\frac{v_{T_0,i}^2}{v_{T_0,1}^2} \leq \left(\frac{2R}{\Delta}\right)^{2T_0} \frac{\sum_{i=2}^d v_{0,i}^2}{v_{0,1}^2} \leq \left(\frac{2R}{\Delta}\right)^{2T} \frac{\sum_{i=2}^d v_{0,i}^2}{v_{0,1}^2},$$

completing the proof. ■

If $\|v_t\| = \gamma_1$, then $v_{t',1} = 0$ for all $t' > t$, and, if $\|v_t\|$ is very close to γ_1, the first component of v_{t+1} could be small enough that it takes a long time to recover. Lemma 7 establishes that this is unlikely.

Lemma 7 *Fix a constant A and $R > 0$. For all $\delta > 0$, if $v_0 \in \mathbb{R}^d$ is chosen randomly from a distribution such that $\mathbf{Pr}[\|v_0\| > R] \leq \delta$, and whose density is bounded above by A, then, with probability $1 - 2\delta$, for all t, $|\|v_t\| - \gamma_1| \geq \Delta$, where*

$$\Delta = \frac{\Gamma(d/2)\delta}{4\pi^{d/2}(2\gamma_1)^{d-1}\bar{T}_0 A}\left(\frac{(\eta\lambda_d)^{d+3}\gamma_1^3}{9 \cdot 6^{d+3}R^3}\right)^{\bar{T}_0}$$

and $\bar{T}_0 = \left\lceil \frac{[\log(R/b)]_+}{\eta\lambda_d} \right\rceil$. Thus,

$$\log\frac{1}{\Delta} \leq \frac{[\log(R/b)]_+ \log\left(\frac{9 \cdot 6^{d+3}R^3}{(\eta\lambda_d)^{d+3}\gamma_1^3}\right)}{\eta\lambda_d} + \log\left(\frac{4\pi^{d/2}(2\gamma_1)^{d-1}[\log(R/b)]_+ A}{\Gamma(d/2)\delta\eta\lambda_d}\right).$$

Proof Before delving into the details, here is the outline of the proof. We argue that at every step when $\|v_t\|$ is bigger than γ_1, the density is small, and hence $\|v_{t+1}\|$ is unlikely to fall in the interval $[\gamma_1 - \Delta, \gamma_1 + \Delta]$. We consider all steps until $\|v_t\| < \gamma_1 + \epsilon$, where ϵ is larger than Δ and is chosen so that, if $\|v_t\| < \gamma_1 + \epsilon$, then $\|v_{t+1}\|$ drops below $\gamma_1 - \epsilon \leq \gamma_1 - \Delta$. We choose $\epsilon = \eta\lambda_d\gamma_1/(2 - \eta\lambda_d)$ for this purpose: the proof of the previous lemma shows that $\|v_{t+1}\| \leq (1 - \eta\lambda_d)\|v_t\|$, and with our choice of ϵ, $\|v_t\| < \gamma_1 + \epsilon$ implies $\|v_{t+1}\| < (1 - \eta\lambda_d)(\gamma_1 + \epsilon) < \gamma_1 - \epsilon$. We compute an upper bound on the factor by which the density increases at each step when $\|v_t\| \geq \gamma_1 + \epsilon$. Lemma 6 shows that there cannot be many such steps.

Let f denote the mapping from v_t to v_{t+1} whose domain is $\{v : \|v\| \geq \gamma_1 + \epsilon\}$, so that if we define $G = \mathrm{diag}(1 - \eta\lambda_1, \ldots, 1 - \eta\lambda_d)$ and $H = \mathrm{diag}((1 - \eta\lambda_1)\gamma_1, \ldots, (1 - \eta\lambda_d)\gamma_d)$, then we can write

$$f(v) = Gv + H\frac{v}{\|v\|}.$$

If μ_t is the density of v_t and f is invertible, and we denote the Jacobian of f^{-1} by ∇f^{-1}, then the density μ_{t+1} of v_{t+1} is the pushforward measure

$$\mu_{t+1}(x) = \mu_t(f^{-1}(x))\left|(\nabla f^{-1})(x)\right|.$$

To see that f is indeed invertible, we write $v = r\hat{v}$, for $r > 0$ and $\|\hat{v}\| = 1$, and $y = f(v)$. Then

$$y = Gv - H\hat{v} = (rG - H)\,\hat{v}.$$

To see that $rG - H$ is invertible, note that

$$\begin{aligned} rG - H &= \mathrm{diag}\left((1 - \eta\lambda_1)r - (1 - \eta\lambda_1)\gamma_1, ..., (1 - \eta\lambda_d)r - (1 - \eta\lambda_d)\gamma_d\right) \\ &= \mathrm{diag}\left((1 - \eta\lambda_1)(r - \gamma_1), ..., (1 - \eta\lambda_d)(r - \gamma_d)\right), \end{aligned}$$

and each entry is positive because $r = \|v_t\| > \gamma_1 \geq \gamma_i$. This means r is the unique solution to $y^\top(rG - H)^{-2}y = 1$ and $\hat{v} = (rG - H)^{-1}y$, and then

$$v = r\hat{v} = r\,(rG - H)^{-1}y = (G - H/r)^{-1}y.$$

To compute the Jacobian of f^{-1}, let's compute dr/dy_j by differentiating the equation defining r. Adopting the shorthand $g_i = G_{ii}$ and $h_i = H_{ii}$, we have

$$\begin{aligned}
& \sum_{i=1}^{d} \frac{d}{dy_j} \frac{y_i^2}{(rg_i - h_i)^2} = 0 \\
\Leftrightarrow \quad & \frac{2y_j}{(rg_j - h_j)^2} - \frac{dr}{dy_j} \sum_{i=1}^{d} \frac{y_i^2 g_i}{2(rg_i - h_i)^3} = 0 \\
\Leftrightarrow \quad & \frac{dr}{dy_j} = \frac{2y_j}{(rg_j - h_j)^2 \sum_{i=1}^{d} \frac{y_i^2 g_i}{2(rg_i - h_i)^3}}.
\end{aligned}$$

Next, we use $v_i = y_i/(g_i - h_i/r)$ to obtain the i, j entry of the Jacobian of f^{-1}:

$$\begin{aligned}
\frac{dv_i}{dy_j} &= \frac{\delta_{ij}}{g_i - h_i/r} + \frac{v_i h_i}{(g_i - h_i/r)^2 r^2} \frac{dr}{dy_j} \\
&= \frac{\delta_{ij}}{g_i - h_i/r} + \frac{2y_i h_i y_j}{(g_i - h_i/r)^2 r^2 (rg_j - h_j)^2 \sum_{k=1}^{d} \frac{v_k^2 g_k}{2(rg_k - h_k)^3}}.
\end{aligned}$$

Assembling these partial derivatives into the Jacobian ∇f^{-1} yields the sum of an invertible diagonal matrix and a rank one matrix. We can use the fact that $\det(A + ab^\top) = \det(A)(1 + b^\top A^{-1} a)$ to get an explicit expression:

$$\det\left(\nabla f^{-1}\right) = \frac{1 + \sum_{i=1}^{d} \frac{2y_i^2 h_i}{(g_i - h_i/r) r^2 (rg_i - h_i)^2 \sum_{k=1}^{d} \frac{y_k^2 g_k}{2(rg_k - h_k)^3}}}{\prod_{i=1}^{d} (g_i - h_i/r)}. \tag{6}$$

Since $r \geq \gamma_1 + \epsilon$, $r \leq \|v_0\|$ and with probability at least $1 - \delta$, $\|v_0\| \leq R$, we have

$$\begin{aligned}
|g_i - h_i/r| &= (1 - \eta\lambda_i)\left(1 - \frac{\gamma_i}{r}\right) \\
&\geq (1 - \eta\lambda_i)\left(1 - \frac{\gamma_i}{\gamma_1 + \epsilon}\right) \\
&\geq (1 - \eta\lambda_1)\left(1 - \frac{\gamma_1}{\gamma_1 + \epsilon}\right) \\
&= (1 - \eta\lambda_1)\left(1 - \frac{1}{1 + \epsilon/\gamma_1}\right) \\
&= (1 - \eta\lambda_1)\left(1 - \frac{1}{1 + \frac{\eta\lambda_d}{2 - \eta\lambda_d}}\right).
\end{aligned}$$

Recalling that $\eta\lambda_d < \eta\lambda_1 \leq 1/2$, this gives

$$|g_i - h_i/r| \geq \frac{\eta\lambda_d}{6}.$$

Defining $B = \frac{\eta\lambda_d}{6}$, substituting into (6), we get

$$\begin{aligned}
\left|\det\left(\nabla f^{-1}\right)\right| &\leq \frac{1}{B^d} + \frac{\|y\|^2 2(1-\eta\lambda_1)\gamma_1}{B^d B r^4 B^2 \sum_{k=1}^d \frac{y_k^2 g_k}{2(rg_k - h_k)^3}} \\
&\leq \frac{1}{B^d} + \frac{2}{B^{d+3}\gamma_1^3 \min_{k=1}^d \frac{g_k}{2(rg_k-h_k)^3}} \\
&\leq \frac{1}{B^d} + \frac{4r^3(1-\eta\lambda_d)^3}{B^{d+3}\gamma_1^3(1-\eta\lambda_1)} \\
&\leq \frac{1}{B^d} + \frac{8R^3}{B^{d+3}\gamma_1^3} \\
&\leq \frac{9R^3}{B^{d+3}\gamma_1^3}.
\end{aligned}$$

Since the density of the initial v_0 is upper bounded by A and Lemma 6 shows that $\|v_{T_0}\| < b < \gamma_1 + \epsilon$, for all $t \leq T_0$ (with T_0 as defined in that lemma), the density in the ring $\{v : \gamma_1 - \Delta \leq \|v\| \leq \gamma_1 + \Delta\}$ is no more than

$$\bar{A} := \left(\frac{9 \cdot 6^{d+3} R^3}{(\eta\lambda_d)^{d+3}\gamma_1^3}\right)^{T_0} A.$$

This implies that for all t,

$$\begin{aligned}
\mathbf{Pr}[\gamma_1 - \Delta \leq \|v_t\| \leq \gamma_1 + \Delta] &\leq 2\Delta S_{d-1}(\gamma_1 + \Delta)\bar{A} \\
&= 2\Delta \frac{2\pi^{d/2}}{\Gamma(d/2)}(\gamma_1 + \Delta)^{d-1}\bar{A} \\
&\leq \Delta \frac{4\pi^{d/2}}{\Gamma(d/2)}(2\gamma_1)^{d-1}\bar{A} \\
&\leq \frac{\delta}{T_0}
\end{aligned}$$

where $S_{d-1}(r)$ is the surface area of the $(d-1)$-sphere of radius r in $\mathbb{R}^d$. Since there are at most T_0 iterations for which $\|v_t\| \geq b$, there are at most T_0 steps for which $\|v_t\| \geq \gamma_1 + \epsilon$. Clearly, $T_0 \leq \bar{T}_0$, which completes the proof. ■

4.3 SAM as gradient descent

The analysis of SAM is complicated by the presence of the $\|\Lambda w_t\|$ term, which couples all d components of the recurrence. Lemma 10 below shows that if we incorporate an alternating sign, we can view SAM as a gradient descent update based on a non-convex objective function J, defined in the following proposition.

Proposition 8 *The function*

$$J(u) = \frac{1}{2}u^\top C u - \|\Lambda u\| = \frac{1}{2}\sum_{i=1}^d \frac{\lambda_i^2 u_i^2}{\beta_i} - \sqrt{\sum_{i=1}^d \lambda_i^2 u_i^2}$$

with

$$C = \frac{1}{\eta\rho}(2I - \eta\Lambda) = \operatorname{diag}\left(\frac{\lambda_1^2}{\beta_1}, \ldots, \frac{\lambda_d^2}{\beta_d}\right)$$

has derivatives

$$\nabla J(u) = Cu - \frac{\Lambda^2 u}{\|\Lambda u\|},$$
$$\nabla^2 J(u) = C - \frac{1}{\|\Lambda u\|}\Lambda P_\perp \Lambda$$

where $P_\perp = I - \Lambda u u^\top \Lambda / \|\Lambda u\|^2$ is the projection on to the subspace orthogonal to Λu. Further, $\nabla J(u) = 0$ iff for some $1 \le i \le d$, $\|u\| = \beta_i/\lambda_i$ and $u \in \operatorname{span}\{e_j : \lambda_j = \lambda_i\}$.

Also, for unit norm $\hat{u}$ satisfying $\nabla J\left(\frac{\beta_i}{\lambda_i}\hat{u}\right) = 0$,

$$\nabla^2 J\left(\frac{\beta_i}{\lambda_i}\hat{u}\right) = \Lambda^2\left(\sum_{j:\beta_j \neq \beta_i}\left(\frac{1}{\beta_j} - \frac{1}{\beta_i}\right)e_j e_j^\top + \frac{1}{\beta_i}e_i e_i^\top\right),$$

which has $|\{j : \beta_j < \beta_i\}| + 1$ positive eigenvalues, $|\{j : \beta_j > \beta_i\}|$ negative eigenvalues, and $|\{j : \beta_j = \beta_i\}| - 1$ zero eigenvalues.

Remark 9 *Although J is not convex, it is well-behaved (see Figure 2). In particular, the set of all stationary points with only non-negative eigenvalues is*

$$M = \left\{u \in \mathbb{R}^d : \|u\| = \frac{\beta_1}{\lambda_1},\, u \in \operatorname{span}\{e_j : \lambda_j = \lambda_1\}\right\},$$

and this is the set of global minima. There are no other local minima, since at all other stationary points the Hessian has a negative eigenvalue. It is easy to see that all $u \in M$ have $J(u) = -\beta_1/2$. And, for example, if $\lambda_1 > \lambda_2$, then $M = \left\{-\frac{\beta_1}{\lambda_1}e_1, \frac{\beta_1}{\lambda_1}e_1\right\}$, and at all other stationary points the Hessian has at least one negative eigenvalue no larger than $1/\beta_1 - 1/\beta_2 < 0$.

Proof We have

$$\nabla J(u) = \nabla\left(\frac{u^\top(I - \eta\Lambda/2)u}{\eta\rho} - \sqrt{u^\top \Lambda^2 u}\right)$$
$$= \frac{2I - \eta\Lambda}{\eta\rho}u - \frac{\Lambda^2 u}{\|\Lambda u\|},$$

and

$$\nabla^2 J(u) = \frac{2I - \eta\Lambda}{\eta\rho} - \frac{\Lambda^2}{\|\Lambda u\|} + \frac{\Lambda^2 u u^\top \Lambda^2}{\|\Lambda u\|^3}$$
$$= \frac{2I - \eta\Lambda}{\eta\rho} - \frac{1}{\|\Lambda u\|}\Lambda\left(I - \frac{\Lambda u u^\top \Lambda}{\|\Lambda u\|^2}\right)\Lambda.$$

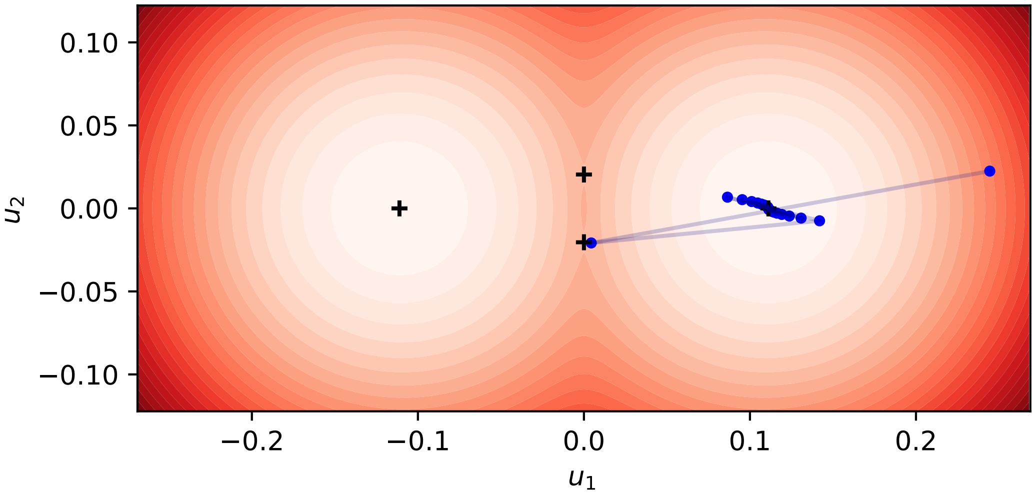


Figure 2: A heat map of the function J defined in Proposition 8 in the case that $d = 2$, $\eta = 1/5$, $\rho = 1$, $\lambda_1 = 1$ and $\lambda_2 = 1/5$, along with the iterates $u_0, u_1, ..., u_{25}$ from Lemma 10 when $w_0 = (0.244, 0.0224)$. The black pluses mark the stationary points.

Now, if u satisfies $\nabla J(u) = 0$, then

$$\begin{aligned} \frac{(2I - \eta\Lambda)u}{\eta\rho} &= \frac{\Lambda^2 u}{\|\Lambda u\|} \\ \Rightarrow \left(\frac{\Lambda^{-1}}{||\Lambda u||}\right) \frac{(2I - \eta\Lambda)u}{\eta\rho} &= \frac{\Lambda u}{\|\Lambda u\|^2} \\ \Rightarrow \left(\frac{\Lambda^{-1}}{||\Lambda u||}\right) \frac{(2I - \eta\Lambda)\Lambda^{-1}\Lambda u}{\eta\rho} &= \frac{\Lambda u}{\|\Lambda u\|^2} \\ \Rightarrow \left(\frac{\Lambda^{-2}(2I - \eta\Lambda)}{\eta\rho}\right) \frac{\Lambda u}{\|\Lambda u\|} &= \frac{\Lambda u}{\|\Lambda u\|^2}, \end{aligned}$$

that is, $\Lambda u/\|\Lambda u\|$ is an eigenvector of $\frac{\Lambda^{-2}(2I-\eta\Lambda)}{\eta\rho} = \mathrm{diag}(1/\beta_1, \ldots, 1/\beta_d)$ with eigenvalue $1/\|\Lambda u\|$. Consider one such stationary point: $\zeta_i e_i$, for some $i \in [d]$ and $\zeta_i \in \mathbb{R}$. We have

$$\frac{(2 - \eta\lambda_i)\zeta_i}{\eta\rho} = \frac{\lambda_i^2 \zeta_i}{\lambda_i |\zeta_i|}$$

which implies

$$|\zeta_i| = \frac{\eta\rho\lambda_i}{2 - \eta\lambda_i} = \frac{\beta_i}{\lambda_i}.$$

Nearly exactly the same reasoning implies that $\|u\| = \frac{\beta_i}{\lambda_i}$ for all stationary points u whose eigenvalues are the same as e_i. For such a stationary point,

$$\begin{aligned}\nabla^2 J(u) &= \frac{2I - \eta\Lambda}{\eta\rho} - \frac{1}{\beta_i}\Lambda\left(I - e_i e_i^\top\right)\Lambda \\ &= \Lambda^2\left(\operatorname{diag}(1/\beta_1, \ldots, 1/\beta_d) - \frac{1}{\beta_i}\left(I - e_i e_i^\top\right)\right) \\ &= \Lambda^2\left(\sum_{j:\beta_j \neq \beta_i}\left(\frac{1}{\beta_j} - \frac{1}{\beta_i}\right)e_j e_j^\top + \frac{1}{\beta_i}e_i e_i^\top\right).\end{aligned}$$

The counts of eigenvalues of different signs follow from this and the ordering $\beta_1 \geq \cdots \geq \beta_d$ (Lemma 5, part 4). ■

The following lemma shows that SAM can be viewed as gradient descent on the objective J. Note that the lemma does not require that $\lambda_1 > \lambda_2$.

Lemma 10 *For $u_t := (-1)^t w_t$, if $\|w_t\| > 0$ for all t, the iteration*

$$w_{t+1,i} = -\eta\rho\lambda_i^2\frac{w_{t,i}}{\|\Lambda w_t\|} + (1 - \eta\lambda_i)w_{t,i}$$

for $i = 1, \ldots, d$ is equivalent to

$$u_{t+1} = u_t - \eta\rho\nabla J(u_t),$$

where J is defined in Proposition 8. Furthermore,

$$J(u_{t+1}) - J(u_t) \leq -\frac{1}{2\rho}\sum_{i=1}^{d} u_{t,i}^2\left(1 - \frac{\beta_i}{\|\Lambda u\|}\right)^2(2 - \eta\lambda_i)^2\lambda_i.$$

Proof From (3),

$$\begin{aligned}u_{t+1} &= (-1)^{t+1}w_{t+1} \\ &= (-1)^{t+1}\left(I - \eta\Lambda - \frac{\eta\rho}{\|\Lambda w_t\|}\Lambda^2\right)w_t \\ &= \eta\rho\Lambda^2\frac{u_t}{\|\Lambda u_t\|} - (I - \eta\Lambda)u_t \\ &= u_t - \eta\rho\left(\left(\frac{2I - \eta\Lambda}{\eta\rho}\right)u_t - \Lambda^2\frac{u_t}{\|\Lambda u_t\|}\right) \\ &= u_t - \eta\rho\nabla\left(\frac{u_t^\top(2I - \eta\Lambda)u_t}{2\eta\rho} - \|\Lambda u_t\|\right) \\ &= u_t - \eta\rho\nabla J(u_t).\end{aligned}$$

Applying the fundamental theorem of calculus twice and using the fact that $\nabla^2 J(u) \preceq (2I - \eta\Lambda)/(\eta\rho)$,

$$\begin{aligned}
&J(u_{t+1}) - J(u_t) \\
&= (u_{t+1} - u_t)^\top \int_0^1 \nabla J(u_t + h(u_{t+1} - u_t))\, dh \\
&= (u_{t+1} - u_t)^\top \int_0^1 \left(\nabla J(u_t) + h\left(\int_0^1 \nabla^2 J(u_t + xh(u_{t+1} - u_t))\, dx\right)(u_{t+1} - u_t)\right) dh \\
&\le \nabla J(u_t)^\top (u_{t+1} - u_t) + \frac{1}{2}(u_{t+1} - u_t)^\top \frac{2I - \eta\Lambda}{\eta\rho}(u_{t+1} - u_t) \\
&= -\eta\rho \nabla J(u_t)^\top \nabla J(u_t) + \eta\rho \nabla J(u_t)^\top \left(I - \frac{\eta\Lambda}{2}\right)\nabla J(u_t) \\
&= -\eta\rho u_t^\top \left(\frac{2I - \eta\Lambda}{\eta\rho} - \frac{\Lambda^2}{\|\Lambda u\|}\right)^2 u_t \\
&\qquad\qquad + \eta\rho u_t^\top \left(\frac{2I - \eta\Lambda}{\eta\rho} - \frac{\Lambda^2}{\|\Lambda u\|}\right)\left(I - \frac{\eta\Lambda}{2}\right)\left(\frac{2I - \eta\Lambda}{\eta\rho} - \frac{\Lambda^2}{\|\Lambda u\|}\right) u_t \\
&= -\eta\rho u_t^\top \left(\frac{2I - \eta\Lambda}{\eta\rho} - \frac{\Lambda^2}{\|\Lambda u\|}\right)\frac{\eta\Lambda}{2}\left(\frac{2I - \eta\Lambda}{\eta\rho} - \frac{\Lambda^2}{\|\Lambda u\|}\right) u_t \\
&= -\eta\rho \sum_{i=1}^d u_{t,i}^2 \left(\frac{2 - \eta\lambda_i}{\eta\rho} - \frac{\lambda_i^2}{\|\Lambda u\|}\right)^2 \frac{\eta\lambda_i}{2} \\
&= -\frac{1}{2\rho}\sum_{i=1}^d u_{t,i}^2 \left(1 - \frac{\eta\rho\lambda_i^2}{(2 - \eta\lambda_i)\|\Lambda u\|}\right)^2 (2 - \eta\lambda_i)^2 \lambda_i \\
&= -\frac{1}{2\rho}\sum_{i=1}^d u_{t,i}^2 \left(1 - \frac{\beta_i}{\|\Lambda u\|}\right)^2 (2 - \eta\lambda_i)^2 \lambda_i,
\end{aligned}$$

where $\beta_i = \eta\rho\lambda_i^2/(2 - \eta\lambda_i)$ as before. ■

The following lemma shows that SAM cannot spend too much time with $\|v_t\|$ large, because J is non-increasing and it decreases a lot when $\|v_t\|$ is large. Lemma 5 part 2 shows that the norm of v_t decreases when the norm is larger than β_1, and the lemma shows in particular that the norm cannot stay much larger than β_1.

Lemma 11 *For $\epsilon > 0$, and $\|v_{T_0}\| \le b$,*

$$\left|\left\{t \ge T_0 : \|v_t\| \ge (1+\epsilon)\beta_1\right\}\right| \le \frac{2}{\eta\epsilon^2\lambda_1\beta_1}\left(\max_{\|\Lambda w\| \le b, s \in \{-1,1\}} J(sw) - \min_u J(u)\right) \le \frac{3\beta_1}{\eta\epsilon^2\lambda_1\beta_d}.$$

Proof From Lemma 5 part 4, $\beta_i \leq \beta_1$, and the definition of β_i implies that $\lambda_i/\beta_i \geq \lambda_1/\beta_1$. Thus, whenever $\|v_t\| \geq (1+\epsilon)\beta_1$, recalling that $\eta\lambda_1 < 1$, Lemma 10 implies

$$\begin{aligned}
J(u_{t+1}) - J(u_t) &\leq -\frac{1}{2\rho}\sum_{i=1}^{d} u_{t,i}^2 \left(1 - \frac{\beta_i}{\|\Lambda u_t\|}\right)^2 (2-\eta\lambda_i)^2\lambda_i \\
&= -\frac{1}{2\rho}\sum_{i=1}^{d} \left(\frac{v_{t,i}}{\lambda_{t,i}}\right)^2 \left(1 - \frac{\beta_i}{\|v_t\|}\right)^2 (2-\eta\lambda_i)^2\lambda_i \\
&= -\frac{\eta}{2}\sum_{i=1}^{d} v_{t,i}^2 \left(1 - \frac{\beta_i}{\|v_t\|}\right)^2 \frac{(2-\eta\lambda_i)\lambda_i}{\beta_i} \\
&\leq -\frac{\eta}{2}\sum_{i=1}^{d} v_{t,i}^2 \left(1 - \frac{\beta_i}{\|v_t\|}\right)^2 \frac{\lambda_i}{\beta_i} \\
&\leq -\frac{\eta}{2}\sum_{i=1}^{d} v_{t,i}^2 \left(1 - \frac{\beta_i}{(1+\epsilon)\beta_1}\right)^2 \frac{\lambda_i}{\beta_i} \\
&\leq -\frac{\eta}{2}\sum_{i=1}^{d} v_{t,i}^2 \left(1 - \frac{\beta_1}{(1+\epsilon)\beta_1}\right)^2 \frac{\lambda_1}{\beta_1} \\
&= -\eta\left(1 - \frac{1}{1+\epsilon}\right)^2 \frac{\lambda_1}{2\beta_1}\|v_t\|^2 \\
&\leq -\eta\left(\frac{\epsilon}{1+\epsilon}\right)^2 \frac{\lambda_1}{2\beta_1}(1+\epsilon)^2\beta_1^2 \\
&= -\frac{\eta\epsilon^2\lambda_1\beta_1}{2},
\end{aligned}$$

and since J is always nonincreasing, this means there can be no more than

$$\frac{2}{\eta\epsilon^2\lambda_1\beta_1}\left(\max_{w\in\mathbb{R}^d, s\in\{-1,1\}:\|\Lambda w\|\leq b} J(sw) - \min_u J(u)\right)$$

iterations like this.

For the last inequality, we have

$$\begin{aligned}
\max_{\|\Lambda w\| \le b, s} J(sw) &= \max_{0 \le z \le b} \max_{\|\Lambda w\| = z} \left(\frac{1}{2} w^\top C w - z \right) \\
&= \max_{0 \le z \le b} \max_{\|v\| = z} \left(\frac{1}{2} v^\top \Lambda^{-1} \mathrm{diag}\left(\frac{\lambda_1^2}{\beta_1}, \ldots, \frac{\lambda_d^2}{\beta_d} \right) \Lambda^{-1} v - z \right) \\
&= \max_{0 \le z \le b} \max_{\|v\| = z} \left(\frac{1}{2} v^\top \mathrm{diag}\left(\frac{1}{\beta_1}, \ldots, \frac{1}{\beta_d} \right) v - z \right) \\
&= \max_{0 \le z \le b} \left(\frac{z^2}{2\beta_d} - z \right) \\
&= \frac{b^2}{2\beta_d} - b \\
&\le \frac{2\beta_1^2}{\beta_d},
\end{aligned}$$

since $b \le 2\beta_1$.

Since $\min_u J(u) = -\beta_1/2$, we have

$$\max_{\|u\| \le b} J(u) - \min_u J(u) \le \frac{2\beta_1^2}{\beta_d} + \frac{\beta_1}{2} \le \frac{3\beta_1^2}{2\beta_d},$$

since $\beta_d \le \beta_1$. ■

4.4 Avoiding non-minimal stationary points

Lemma 10 shows that the set of global minima of J is a sphere of radius β_1/λ_1 in the subspace spanned by the e_j with $\lambda_j = \lambda_1$. To simplify notation, we assume that $\lambda_1 > \lambda_2$, so that this subspace is in the e_1 direction. Then to ensure that J decreases to a global minimum, it suffices to keep $|\lambda_1 w_{t,1}| = |v_{t,1}|$ away from zero and $\|v_t\| \ne \beta_1$. The following quantity measures the extent to which v_t still has "energy" in components other than the first.

Definition 12 *Define* $\delta_t = 1 - \frac{|v_{t,1}|}{\|v_t\|}$.

Lemma 13 *We have*

$$\delta_t \le \frac{1}{2} \frac{\sum_{i=2}^d v_{t,i}^2}{v_{t,1}^2}$$

whenever this bound is most $1/2$.

Proof We have

$$\begin{aligned}
\delta_t &= 1 - \frac{|v_{t,1}|}{\|v_t\|} \\
&= 1 - \frac{1}{\sqrt{1 + \sum_{i=2}^d v_{t,i}^2 / v_{t,1}^2}} \\
&\leq \frac{1}{2} \frac{\sum_{i=2}^d v_{t,i}^2}{v_{t,1}^2}
\end{aligned}$$

since, for all $0 \leq \alpha \leq 1$, we have $1 - 1/\sqrt{1+\alpha} \leq \alpha/2$. Indeed, this inequality is equivalent to

$$\begin{aligned}
1 &\geq (1+\alpha)\left(1 - \frac{\alpha}{2}\right)^2 \\
&= (1+\alpha)\left(1 - \alpha + \frac{\alpha^2}{4}\right) \\
&= 1 - \alpha^2 + \frac{\alpha^2}{4} + \frac{\alpha^3}{4} \\
&= 1 - \frac{\alpha^2(3-\alpha)}{4}.
\end{aligned}$$

■

Lemma 5 part 3 shows that the first component increases relative to the other components when $\|v_t\| < \alpha_d$. But as long as $\lambda_d > 0$, part 5 shows that $\alpha_d > \beta_1$, and in that case Lemma 11 implies that $\|v_t\|$ does not spend too much time above α_d. Our assumption that $\lambda_1 > \lambda_2$ ensures that the first component of v_t increases in magnitude relative to all the other components; otherwise, the equations describing the evolution of the first and second components are identical. The key constant depends on both λ_d and the gap between λ_1 and λ_2.

Lemma 14 *Define*

$$\mu = \min\left\{\eta\lambda_d, \frac{\lambda_1^2}{\lambda_2^2} - 1\right\}.$$

If $v_{t,1}^2 > 0$, the following two statements are equivalent:

$$\begin{aligned}
\frac{v_{t+1,i}^2}{v_{t+1,1}^2} &< \frac{1}{(1+\mu)^2} \frac{v_{t,i}^2}{v_{t,1}^2} \qquad \forall i \in \{2, \ldots, d\}, \\
\frac{\|v_t\|}{\beta_1} &< \frac{2 - \eta\lambda_1}{2 - \eta\lambda_1 - (\eta\lambda_d - \mu) - \eta\lambda_d\mu} \left(1 + (1+\mu)\frac{\lambda_d^2}{\lambda_1^2}\right).
\end{aligned}$$

Thus, if $v_{t,1}^2 > 0$ for all t,

$$\left|\left\{t : \|v_t\| \leq b \text{ and for some } i \in \{2, \ldots, d\}, \; \frac{v_{t+1,i}^2}{v_{t+1,1}^2} \geq \frac{1}{(1+\mu)^2} \frac{v_{t,i}^2}{v_{t,1}^2}\right\}\right| \leq T_1,$$

where

$$T_1 = \frac{3\beta_1\lambda_1^3}{\eta\beta_d\lambda_d^4}.$$

Furthermore, if T_0 *is such that* $\|v_{T_0}\| \le b$*, then*

$$\delta_{T+1} \le \frac{1}{2}\left(\frac{1}{1+\mu}\right)^{2(T-T_1)}\left(\frac{2}{\eta\lambda_1}\right)^{2T_1}\frac{\sum_{i=2}^d v_{T_0,i}^2}{v_{T_0,1}^2},$$

provided that T *is large enough that this upper bound is less than* $1/2$*.*

Thus, for all $\epsilon < 1/2$*,* $\delta_{T+1} \le \epsilon$ *provided*

$$T \ge \frac{2}{\mu}\left(T_1 \log\left(\frac{4}{\eta\lambda_1}\right) + \frac{1}{2}\log\left(\frac{\sum_{i=2}^d v_{T_0}^2}{2\epsilon v_{T_0,1}^2}\right)\right).$$

Proof For the equivalence, notice that the evolution of v_t implies that

$$\frac{v_{t+1,i}^2}{v_{t+1,1}^2} < \frac{1}{(1+\mu)^2}\frac{v_{t,i}^2}{v_{t,1}^2}$$

if and only if

$$(1-\eta\lambda_1)^2(\|v_t\| - \gamma_1)^2 > (1+\mu)^2(1-\eta\lambda_i)^2(\|v_t\| - \gamma_i)^2. \tag{7}$$

We can apply Lemma 4, because $0 < 1-\eta\lambda_1 < (1+\mu)(1-\eta\lambda_i)$, $\gamma_1 > \gamma_i$, and

$$\begin{aligned} & (1-\eta\lambda_1)\gamma_1 \ge (1+\mu)(1-\eta\lambda_i)\gamma_i \\ \Leftrightarrow \quad & \lambda_1^2 \ge (1+\mu)\lambda_i^2 \\ \Leftrightarrow \quad & \frac{\lambda_1^2}{\lambda_i^2} - 1 \ge \mu, \end{aligned}$$

which follows from the definition of μ. Lemma 4 implies that when $\|v_t\| > 0$, (7) is equivalent to

$$\|v_t\| < \frac{(1-\eta\lambda_1)\gamma_1 + (1+\mu)(1-\eta\lambda_i)\gamma_i}{1-\eta\lambda_1 + (1+\mu)(1-\eta\lambda_i)}.$$

Because the right hand side is a convex combination of γ_1 and γ_i, because $\gamma_d \le \cdots \le \gamma_2$, and because the convex coefficients are also ordered $(1-\eta\lambda_2 \le \cdots \le 1-\eta\lambda_d)$, these inequalities for all i are implied by the corresponding inequality for $i = d$, which is

$$\begin{aligned} & \|v_t\| < \frac{(1-\eta\lambda_1)\gamma_1 + (1+\mu)(1-\eta\lambda_d)\gamma_d}{1-\eta\lambda_1 + (1+\mu)(1-\eta\lambda_d)} \\ \Leftrightarrow \quad & \frac{\|v_t\|}{\beta_1} < \frac{2-\eta\lambda_1}{(1-\eta\lambda_1)\gamma_1}\left(\frac{(1-\eta\lambda_1)\gamma_1 + (1+\mu)(1-\eta\lambda_d)\gamma_d}{1-\eta\lambda_1 + (1+\mu)(1-\eta\lambda_d)}\right) \\ & \quad = \frac{2-\eta\lambda_1}{2-\eta\lambda_1 - (\eta\lambda_d - \mu) - \eta\lambda_d\mu}\left(1 + \frac{(1+\mu)(1-\eta\lambda_d)\gamma_d}{(1-\eta\lambda_1)\gamma_1}\right) \\ & \quad = \frac{2-\eta\lambda_1}{2-\eta\lambda_1 - (\eta\lambda_d - \mu) - \eta\lambda_d\mu}\left(1 + (1+\mu)\frac{\lambda_d^2}{\lambda_1^2}\right). \end{aligned}$$

This proves the first part of the lemma.

Hence, for each iteration when, for some $2 \le i \le d$,

$$\frac{v_{t+1,i}^2}{v_{t+1,1}^2} \ge \frac{1}{(1+\mu)^2}\frac{v_{t,i}^2}{v_{t,1}^2},$$

we must have

$$\frac{\|v_t\|}{\beta_1} \ge \frac{2-\eta\lambda_1}{2-\eta\lambda_1-(\eta\lambda_d-\mu)-\eta\lambda_d\mu}\left(1+(1+\mu)\frac{\lambda_d^2}{\lambda_1^2}\right) > 1+(1+\mu)\frac{\lambda_d^2}{\lambda_1^2},$$

where the inequality follows from $0 < \mu \le \eta\lambda_d$ and $\eta < 1/\lambda_1$. Lemma 11 implies that the number of these iterations for which we also have $\|v_t\| \le b$ is no more than

$$\frac{3\beta_1}{\eta\left((1+\mu)\frac{\lambda_d^2}{\lambda_1^2}\right)^2\lambda_1\beta_d} \le \frac{3\beta_1\lambda_1^3}{\eta\beta_d\lambda_d^4} = T_1.$$

For the third part, consider the sequence of steps from $t = T_0$ to $t = T \ge T_1$. There are at least $T - T_1$ steps when

$$\frac{v_{t+1,i}^2}{v_{t+1,1}^2} < \frac{1}{(1+\mu)^2}\frac{v_{t,i}^2}{v_{t,1}^2},$$

and no more than T_1 steps when this fails, and for those steps we have

$$\begin{aligned}\frac{v_{t+1,i}^2}{v_{t+1,1}^2} &= \frac{(1-\eta\lambda_i)^2(\|v_t\|-\gamma_i)^2}{(1-\eta\lambda_1)^2(\|v_t\|-\gamma_1)^2}\frac{v_{t,i}^2}{v_{t,1}^2}\\ &\le \frac{(1-\eta\lambda_d)^2\gamma_1^2}{(1-\eta\lambda_1)^2(b-\gamma_1)^2}\frac{v_{t,i}^2}{v_{t,1}^2}\\ &= \frac{(1-\eta\lambda_d)^2}{(1-\eta\lambda_1)^2\eta^2\lambda_1^2}\frac{v_{t,i}^2}{v_{t,1}^2}.\end{aligned}$$

(We used the fact that $0 \le \|v_t\| \le b \le \gamma_1$, and so $(\gamma_i - \|v_t\|)^2 \le \gamma_1^2$.) So we have

$$\frac{\sum_{i=2}^d v_{T+1,i}^2}{v_{T+1,1}^2} \le \left(\frac{1}{(1+\mu)^2}\right)^{T-T_1}\left(\frac{(1-\eta\lambda_d)^2}{(1-\eta\lambda_1)^2\eta^2\lambda_1^2}\right)^{T_1}\frac{\sum_{i=2}^d v_{T_0,i}^2}{v_{T_0,1}^2}.$$

Applying Lemma 13

$$\delta_t \le \frac{1}{2}\left(\frac{1}{(1+\mu)^2}\right)^{T-T_1}\left(\frac{(1-\eta\lambda_d)^2}{(1-\eta\lambda_1)^2\eta^2\lambda_1^2}\right)^{T_1}\frac{\sum_{i=2}^d v_{T_0,i}^2}{v_{T_0,1}^2},$$

if this bound is at most 1/2. Solving for T, for all $0 < \epsilon < 1/2$, for all

$$T \ge \frac{1}{\log(1+\mu)}\left(T_1\log\left(\frac{(1+\mu)(1-\eta\lambda_d)}{(1-\eta\lambda_1)\eta\lambda_1}\right)+\frac{1}{2}\log\left(\frac{\sum_{i=2}^d v_{T_0}^2}{2\epsilon v_{T_0,1}^2}\right)\right),$$

we have $\delta_t \le \epsilon$. Noting that $\mu = \min\left\{\eta\lambda_d, \frac{\lambda_1^2}{\lambda_2^2} - 1\right\} \le \eta\lambda_d \le \eta\lambda_1 \le 1/2$ completes the proof. ■

Once the first component of v_t dominates, the recurrence becomes essentially one-dimensional, and its convergence is easier to analyze, as the following lemma shows.

Definition 15 *Let* $s_t = \operatorname{sign}(v_{t,1})$.

Lemma 16 *If* $\|v_t\| > 0$,

$$v_{t+1,1} - (-s_t\beta_1) = -(1-\eta\lambda_1)\left(v_{t,1} - s_t\beta_1 + s_t\gamma_1\delta_t\right).$$

If $0 < ||v_T|| \le b$ *and for all* $t \ge T$, $\delta_t \le \frac{\eta\lambda_1\beta_1}{2}$, *then for all* $t \ge T$,

$$\left|v_{t+1,1} - (-1)^{t+1-T} s_T\beta_1\right| \le (1-\eta\lambda_1)\left(\left|v_{t,1} - (-1)^{t-T}s_T\beta_1\right| + \gamma_1\delta_t\right).$$

Proof From the recurrence for v_t, we have

$$\begin{aligned}
v_{t+1,1} &= (1-\eta\lambda_1)\left(1 - \frac{\gamma_1}{\|v_t\|}\right)v_{t,1} \\
&= (1-\eta\lambda_1)v_{t,1} - (2-\eta\lambda_1)s_t\beta_1\frac{|v_{t,1}|}{\|v_t\|} \\
&= (1-\eta\lambda_1)v_{t,1} - s_t\beta_1\left(1-(2-\eta\lambda_1)\left(1-\frac{|v_{t,1}|}{\|v_t\|}\right)\right) + s_t\beta_1(1-(2-\eta\lambda_1)) \\
&= (1-\eta\lambda_1)(v_{t,1} - s_t\beta_1) - s_t\beta_1\left(1-(2-\eta\lambda_1)\delta_t\right).
\end{aligned}$$

So

$$\begin{aligned}
v_{t+1,1} - (-s_t\beta_1) &= -(1-\eta\lambda_1)\left(v_{t,1} - s_t\beta_1\right) + s_t\beta_1(2-\eta\lambda_1)\delta_t \\
&= -(1-\eta\lambda_1)\left(v_{t,1} - s_t\beta_1\right) + s_t\gamma_1(1-\eta\lambda_1)\delta_t \\
&= -(1-\eta\lambda_1)\left(v_{t,1} - s_t\beta_1 + s_t\gamma_1\delta_t\right), \qquad (8)
\end{aligned}$$

which is the equality in the lemma.

Since $||v_T|| \le b$, Part 5 of Lemma 5 implies that for all $t \ge T$, $||v_t|| \le b$, which in turn implies $0 \le s_t v_{t,1} \le b$. Since $0 \le b/2 < \beta_1$, and $|v_{t,1} - s_t\beta_1| = |s_t v_{t,1} - \beta_1|$, this implies, for all $t \ge T$,

$$|v_{t,1} - s_t\beta_t| \le b - \beta_1 = (1-\eta\lambda_1)\beta_1. \qquad (9)$$

By the triangle inequality for the absolute difference, since $\delta_t \ge 0$,

$$|v_{t+1,1} - (-s_t\beta_1)| \le (1-\eta\lambda_1)\left(|v_{t,1} - s_t\beta_1| + \gamma_1\delta_t\right),$$

which in turn implies

$$\min\left\{|v_{t+1,1} - \beta_1|, |v_{t+1,1} + \beta_1|\right\} \le (1-\eta\lambda_1)\left(|v_{t,1} - s_t\beta_1| + \gamma_1\delta_t\right).$$

Because $\beta_1 > 0$,

$$|v_{t+1,1} - s_{t+1}\beta_1| = \min\left\{|v_{t+1,1} - \beta_1|, |v_{t+1,1} + \beta_1|\right\},$$

so

$$|v_{t+1,1} - s_{t+1}\beta_1| \leq (1 - \eta\lambda_1)\left(|v_{t,1} - s_t\beta_1| + \gamma_1\delta_t\right). \tag{10}$$

It remains to show that, for all $t \geq T$, if $\delta_t \leq \frac{\eta\lambda_1\beta_1}{2}$, then $s_{t+1} = -s_t$.

To see this, assume as a first case that $s_t = 1$. Then (8) implies

$$\begin{aligned} v_{t+1,1} &= -\beta_1 - (1 - \eta\lambda_1)\left(v_{t,1} - \beta_1 + \gamma_1\delta_t\right) \\ &\leq -\beta_1 + (1 - \eta\lambda_1)\left(|v_{t,1} - \beta_1| + \gamma_1\delta_t\right) \\ &\leq -\beta_1 + (1 - \eta\lambda_1)\left((1 - \eta\lambda_1)\beta_1 + \gamma_1\delta_t\right) && \text{(by (9))} \\ &< 0, \end{aligned}$$

since $\delta_t \leq \frac{\eta\lambda_1\beta_1}{2}$, so $s_{t+1} = -1$.

Similarly, if $s_t = -1$, then

$$\begin{aligned} v_{t+1,1} &= \beta_1 - (1 - \eta\lambda_1)\left(v_{t,1} - \beta_1 + \gamma_1\delta_t\right) \\ &\geq \beta_1 - (1 - \eta\lambda_1)\left(|v_{t,1} - (-\beta_1)| + \gamma_1\delta_t\right) \\ &> 0, \end{aligned}$$

so $s_{t+1} = 1$. The last inequality of the lemma then follows by induction. ■

Lemma 17 *If T_0 is the first iteration where $\|v_{T_0}\| \leq b$, then, for all*

$$0 < \epsilon < \min\left\{\sqrt{\frac{\eta\lambda_1\beta_1}{2\gamma_1}}, \frac{\eta\lambda_1}{2\gamma_1}, \frac{1}{b}, \frac{\beta_1}{2}\right\},$$

for

$$T_2 = \frac{2}{\mu}\left(\frac{3\beta_1\lambda_1^3}{\eta\beta_d\lambda_d^4}\log\left(\frac{4}{\eta\lambda_1}\right) + \frac{1}{2}\log\left(\frac{\sum_{i=2}^d v_{T_0}^2}{2\epsilon^2 v_{T_0,1}^2}\right)\right) + \frac{6}{\eta\lambda_1}\ln\left(\frac{1}{\epsilon}\right),$$

for all $t \geq T_2$, we have

$$|v_{t,1} - (-1)^{t-T_2}s_{T_2}\beta_1| \leq \epsilon \text{ and } \delta_t \leq \epsilon^2.$$

Proof The last inequality of Lemma 14 implies that, for

$$t^* \overset{\text{def}}{=} \left\lceil \frac{2}{\mu}\left(\frac{3\beta_1\lambda_1^3}{\eta\beta_d\lambda_d^4}\log\left(\frac{4}{\eta\lambda_1}\right) + \frac{1}{2}\log\left(\frac{\sum_{i=2}^d v_{T_0}^2}{2\epsilon^2 v_{T_0,1}^2}\right)\right)\right\rceil,$$

we have

$$\forall t \geq t^*, \delta_t \leq \epsilon^2. \tag{11}$$

For all $t \geq t^*$, since $\delta_t \leq \epsilon^2 \leq \frac{\eta\lambda_1\beta_1}{2\gamma_1}$, by Lemma 16, we have

$$\left| v_{t+1,1} - (-1)^{t+1-t^*} s_{t^*}\beta_1 \right| \leq (1 - \eta\lambda_1) \left| v_{t,1} - (-1)^{t-t^*} s_{t^*}\beta_1 \right| + \gamma_1\epsilon^2.$$

If $\left| v_{t,1} - (-1)^{t-t^*} s_{t^*}\beta_1 \right| > \epsilon$, this implies

$$\left| v_{t+1,1} - (-1)^{t+1-t^*} s_{t^*}\beta_1 \right| \leq (1 - \eta\lambda_1 + \gamma_1\epsilon) \left| v_{t,1} - (-1)^{t-t^*} s_{t^*}\beta_1 \right|.$$

Since $\epsilon \leq \frac{\eta\lambda_1}{2\gamma_1}$, this yields

$$\left| v_{t+1,1} - (-1)^{t+1-t^*} s_{t^*}\beta_1 \right| \leq \left(1 - \frac{\eta\lambda_1}{2}\right) \left| v_{t,1} - (-1)^{t-t^*} s_{t^*}\beta_1 \right|.$$

Since $\|v_{t^*}\| \leq b$, $|v_{t^*,1}| \leq b$, which, since $\beta_1 \leq b$, implies $|v_{t^*,1} - s_{t^*}\beta_1| \leq b$. Thus, by induction, for all $t \geq t^*$, we have

$$\left| v_{t+1,1} - (-1)^{t+1-t^*} s_{t^*}\beta_1 \right| \leq \left(1 - \frac{\eta\lambda_1}{2}\right)^{t-t^*} b.$$

Thus, if $t \geq T_2 = t^* + \frac{2}{\eta\lambda_1} \ln\left(\frac{b}{\epsilon}\right)$, we get $\left| v_{t+1,1} - (-1)^{t+1-t^*} s_{t^*}\beta_1 \right| \leq \epsilon$. Since $\epsilon < \beta/2$, this implies $s_{t+1} = \text{sign}(v_{t+1,1}) = (-1)^{t+1-t^*} s_{t^*}$. Since, $\epsilon \leq 1/b$, this completes the proof. ■

Lemma 18 *For all* $0 < \epsilon \leq 1$, *if* $|v_{t,1} - (-1)^{t-T_2} s_{T_2}\beta_1| \leq \epsilon$ *and* $\delta_t \leq \epsilon^2$, *then*

$$\|v_t - (-1)^{t-T_2} s_{T_2}\beta_1 e_1\| \leq 2(1 + \beta_1)\epsilon.$$

Proof If $\delta_t \leq \epsilon^2$, then

$$\frac{v_{t,1}^2}{\|v_t\|^2} \geq (1 - \epsilon^2)^2. \tag{12}$$

We have

$$\begin{aligned}
&\|v_t - (-1)^{t-T_2} s_{T_2}\beta_1 e_1\|^2 \\
&= (v_{t,1} - (-1)^{t-T_2} s_{T_2}\beta_1)^2 + \sum_{i>2} v_{t,i}^2 \\
&\leq \epsilon^2 + \sum_{i>2} v_{t,i}^2 \\
&= \epsilon^2 + \|v_t\|^2 - v_{t,1}^2 \\
&\leq \epsilon^2 + \left(\frac{1}{(1-\epsilon^2)^2} - 1\right) v_{t,1}^2 \quad \text{(by (12))} \\
&\leq \epsilon^2(1 + 3v_{t,1}^2) \quad \text{(since } 0 < \epsilon \leq 1) \\
&\leq \epsilon^2(1 + 3(\epsilon + \beta_1)^2)
\end{aligned}$$

since $|v_{t,1} - (-1)^{t-T_2} s_{T_2}\beta_1| \le \epsilon$. Since $\sqrt{1+3(1+x)^2)} \le 2(1+x)$ for all $x > 0$, this completes the proof. ■

Lemma 19 *If T_0 is the first iteration where $\|v_{T_0}\| \le b$, and T_2 is defined as in Lemma 17, then, for all $0 < \epsilon < \min\left\{\sqrt{\frac{2\eta\lambda_1\beta_1}{\gamma_1}}, \frac{\eta\lambda_1}{2\gamma_1}, \frac{2}{b}, \beta_1, 1\right\}$ for any*

$$t \ge \frac{2}{\mu}\left(\frac{3\beta_1\lambda_1^3}{\eta\beta_d\lambda_d^4}\log\left(\frac{4}{\eta\lambda_1}\right) + \frac{1}{2}\log\left(\frac{2(1+\beta_1)^2\sum_{i=2}^d v_{T_0}^2}{\epsilon^2 v_{T_0,1}^2}\right)\right) + \frac{6}{\eta\lambda_1}\ln\left(\frac{2(1+\beta_1)}{\epsilon}\right)$$

we have

$$\|v_t - (-1)^{t-T_2} s_{T_2}\beta_1 e_1\| \le \epsilon.$$

Proof Combine Lemmas 17 and 18. ■

Lemma 20 *For any $s \in \{-1, 1\}$, and any t,*

$$\left\|w_t - \frac{s\beta_1 e_1}{\lambda_1}\right\| \le \frac{\|v_t - s\beta_1 e_1\|}{\lambda_d}.$$

Proof Since $w_t = \Lambda^{-1}v_t$, we have

$$\begin{aligned}\left\|w_t - \frac{s\beta_1 e_1}{\lambda_1}\right\| &= \|\Lambda^{-1}v_t - \Lambda^{-1}s\beta_1 e_1\| \\ &\le \|\Lambda^{-1}\|\|v_t - s\beta_1 e_1\| \\ &= \frac{1}{\lambda_d}\|v_t - s\beta_1 e_1\|\end{aligned}$$

■

4.5 Putting it together

In this subsection, we combine the lemmas proved in earlier subsections to prove Theorem 3. For $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_d)$, our analysis tracks the evolution of $v_t = \nabla\ell(w_t) = \Lambda w_t$.

By assumption, with probability $1-\delta$, $\|w_0\| \le R$ and $w_{0,1}^2 \ge q$. Let us assume from here on that this is the case. This implies $\|v_0\| \le \lambda_1 R$ and $v_{0,1}^2 \ge \lambda_1^2 q$.

Let T_0 be the index of the first iteration that $\|v_t\| \le b$ holds.

Lemmas 6 and 7 imply that, with probability $1-2\delta$, for Δ defined as in Lemma 7, we have

$$\begin{aligned}\log\left(\frac{\sum_{i=2}^d v_{T_0,i}^2}{v_{T_0,1}^2}\right) &\le \frac{2}{\eta\lambda_d}\left[\log\left(\frac{\lambda_1 R}{b}\right)\right]_+ \log\left(\frac{2\lambda_1 R}{\Delta}\right) + \log\left(\frac{\lambda_1^2 R^2}{v_{0,1}^2}\right) \\ &\le \frac{2}{\eta\lambda_d}\left[\log\left(\frac{\lambda_1 R}{b}\right)\right]_+ \log\left(\frac{2\lambda_1 R}{\Delta}\right) + \log\left(\frac{R^2}{q}\right). \qquad (13)\end{aligned}$$

Let us assume for the rest of this proof that this is the case.

Combining (13) with Lemma 19, for all

$$\begin{aligned} t \geq\ & \frac{6\beta_1\lambda_1^3}{\eta\mu\beta_d\lambda_d^4}\log\left(\frac{4}{\eta\lambda_1}\right) \\ & + \frac{1}{\mu}\left(\log\left(\frac{2(1+\beta_1)^2}{\lambda_d^2\epsilon^2}\right) + \frac{2}{\eta\lambda_d}\left[\log\left(\frac{\lambda_1 R}{b}\right)\right]_+ \log\left(\frac{2\lambda_1 R}{\Delta}\right) + \log\left(\frac{R^2}{q}\right)\right) \\ & + \frac{6}{\eta\lambda_1}\ln\left(\frac{2(1+\beta_1)}{\lambda_d\epsilon}\right) \end{aligned}$$

we have

$$\|v_t - (-1)^{t-T_2} s_{T_2}\beta_1 e_1\| \leq \lambda_d\epsilon. \tag{14}$$

Applying Lemma 7 to bound $\log\frac{1}{\Delta}$, we get that

$$\begin{aligned} t \geq\ & \frac{6\beta_1\lambda_1^3}{\eta\mu\beta_d\lambda_d^4}\log\left(\frac{4}{\eta\lambda_1}\right) \\ & + \frac{1}{\mu}\left(\log\left(\frac{2(1+\beta_1)^2}{\lambda_d^2\epsilon^2}\right) + \log\left(\frac{R^2}{q}\right)\right) \\ & + \frac{2}{\eta\lambda_d\mu}\left[\log\left(\frac{\lambda_1 R}{b}\right)\right]_+ \left(\log(2\lambda_1 R) + \frac{[\log(\lambda_1 R/b)]_+ \log\left(\frac{9\cdot 6^{d+3}\lambda_1^3 R^3}{(\eta\lambda_d)^{d+3}\gamma_1^3}\right)}{\eta\lambda_d}\right. \\ & \qquad\qquad \left. + \log\left(\frac{4\pi^{d/2}(2\gamma_1)^{d-1}[\log(\lambda_1 R/b)]_+ A}{\Gamma(d/2)\delta\eta\lambda_d}\right)\right) \\ & + \frac{6}{\eta\lambda_1}\ln\left(\frac{2(1+\beta_1)}{\lambda_d\epsilon}\right) \end{aligned}$$

suffices for (14). Substituting the values of μ, β_1, β_d, γ_1 and b, simplifying and overapproximating, we get that

$$\begin{aligned} t \geq\ & \frac{6\lambda_1^5}{\eta\lambda_d^6\min\left\{\eta\lambda_d, \frac{\lambda_1^2}{\lambda_2^2}-1\right\}}\log\left(\frac{4}{\eta\lambda_1}\right) \\ & + \frac{1}{\min\left\{\eta\lambda_d, \frac{\lambda_1^2}{\lambda_2^2}-1\right\}}\left(\log\left(\frac{4(1+\eta\rho\lambda_1^2)^2}{\lambda_d^2\epsilon^2}\right) + \log\left(\frac{R^2}{q}\right)\right) \\ & + \frac{2\left[\log\left(\frac{R}{\eta\rho\lambda_1}\right)\right]_+}{\eta\lambda_d\min\left\{\eta\lambda_d, \frac{\lambda_1^2}{\lambda_2^2}-1\right\}}\left(\log(2\lambda_1 R) + \frac{\left[\log\left(\frac{R}{\eta\rho\lambda_1}\right)\right]_+ \log\left(\frac{9\cdot 6^{d+3}R^3}{(\eta\lambda_d)^{d+3}(\eta\rho\lambda_1)^3}\right)}{\eta\lambda_d}\right. \\ & \qquad\qquad \left. + \log\left(\frac{4\pi^{d/2}(4\eta\rho\lambda_1^2)^{d-1}\left[\log\left(\frac{R}{\eta\rho\lambda_1}\right)\right]_+ A}{\Gamma(d/2)\delta\eta\lambda_d}\right)\right) \\ & + \frac{6}{\eta\lambda_1}\ln\left(\frac{2(1+\eta\rho\lambda_1^2)}{\lambda_d\epsilon}\right) \end{aligned}$$

suffices.

Applying Lemma 20 completes the proof.

5. Drifting Towards Wide Minima

We have seen that when SAM is applied to a convex quadratic objective, it converges to an oscillation that bounces across the minimum in the direction of greatest curvature. In this section, we consider the behavior of SAM when it is applied to a smooth objective ℓ whose Hessian may vary. Consider a point $w_z \in \mathbb{R}^d$ in a d-dimensional parameter space that is a local minimum of ℓ, $\nabla \ell(w_z) = 0$. For notational convenience, we assume that

$$H := \nabla^2 \ell(w_z) = \mathrm{diag}(\lambda_1, \dots, \lambda_d).$$

In the neighborhood of w_z, the smooth objective ℓ can be approximated locally by the quadratic objective

$$\ell_q(w) = \ell(w_z) + \frac{1}{2}(w - w_z)^\top H (w - w_z).$$

We are particularly interested in the overparameterized setting typical of deep learning, that is, where the dimension of the parameter space exceeds the sample size so that there are many directions in parameter space that do not affect the training objective. Suppose, in particular, that $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_k > \lambda_{k+1} = \cdots = \lambda_d = 0$ for $k > 1$. Then since this quadratic objective does not vary in the $e_{k+1}, \dots, e_d$ directions, for a point w_0 satisfying $e_i^\top (w_0 - w_z) = 0$ for $i = k+1, \dots, d$, if we initialize SAM at w_0 and apply it to the quadratic objective ℓ_q, then it is clear that the condition $e_i^\top (w_t - w_z) = 0$ for $i > k$ continues to hold for all t. Thus, the result above shows that SAM converges to the set

$$\left\{ w_z \pm \frac{\beta_1}{\lambda_1} e_1 \right\}.$$

The following theorem considers SAM's behavior on the smooth objective ℓ at these points. It shows that SAM's gradient steps have a component that maintains the oscillation in the e_1 direction, a second-order component in the downhill direction of the spectral norm of the Hessian, and a third-order component that is small if the third derivative changes slowly. For a symmetric matrix M, $\lambda_{\max}(M)$ denotes the maximum eigenvalue of M. In this section, we write D^i as the symmetric, multilinear, ith-derivative operator and ∇^1 and ∇^2 as the vector and matrix representations of the operators D^1 and D^2 in the canonical basis $e_1, \dots, e_d$.

Theorem 21 *Suppose that ℓ is in C^3, that $D^3\ell$ is B-Lipschitz with respect to the Euclidean norm and the operator norm, and that $w_z \in \mathbb{R}^d$ satisfies $\nabla \ell(w_z) = 0$ and $\nabla^2 \ell(w_z) = \sum_{i=1}^d \lambda_i e_i \otimes e_i$. For $s_t \in \{-1, 1\}$, consider the point*

$$w_t = w_z + \frac{s_t \beta_1}{\lambda_1} e_1 = w_z + \frac{\eta \rho \lambda_1 s_t}{2 - \eta \lambda_1} e_1.$$

Then, if $B\eta\rho \leq 1$, SAM's update on ℓ gives

$$w_{t+1} - w_t = -2\frac{\eta\rho\lambda_1 s_t}{2-\eta\lambda_1}e_1 - \frac{\eta\rho^2}{2}\left(1+\frac{\eta\lambda_1}{2-\eta\lambda_1}\right)^2 \nabla\,\lambda_{\max}(\nabla^2\ell(w_z)) \\ + \eta\rho^2\left(\frac{(1+\eta\lambda_1)^3}{6}\rho + 2(2\lambda_1 + B\rho)\eta\right)B\zeta,$$

where $\|\zeta\| \leq 1$.

Thus, if we define $\epsilon := w_t - w_z$, then for any $\rho \leq c$ and $\eta \leq c\rho$ for some constant c, there are constants c_1 and c_2 that depend on c, B and λ_1 so that

$$w_{t+1} - w_t = -2\epsilon + \|\epsilon\|\rho\left(c_1\nabla\,\lambda_{max}\left(\nabla^2\ell(w_z)\right) + c_2\rho\zeta\right).$$

Proof Let

$$w_u = w_t + \rho\frac{\nabla\ell(w_t)}{\|\nabla\ell(w_t)\|}$$

so that

$$w_{t+1} - w_t = -\eta\nabla\ell(w_u).$$

Let

$$\tilde{w}_u = w_t + s_t\rho e_t = w_z + s_t(\beta_1/\lambda_1 + \rho)e_1.$$

(It may be helpful to think of $\tilde{w}_u$ as what w_u would have been, if SAM used ℓ_q instead of ℓ.) We have

$$w_{t+1} - w_t = -\eta\nabla\ell(\tilde{w}_u) + \eta(\nabla\ell(\tilde{w}_u) - \nabla\ell(w_u)). \tag{15}$$

First, we analyze $\nabla\ell(\tilde{w}_u)$.

The fundamental theorem of calculus implies

$$\begin{aligned}
&D^2\ell(w_z + \epsilon e_1)(\cdot,\cdot) \\
&= D^2\ell(w_z)(\cdot,\cdot) + \int_0^1 D^3\ell(w_z + x\epsilon e_1)(\epsilon e_1,\cdot,\cdot)\,dx \\
&= D^2\ell(w_z)(\cdot,\cdot) + \int_0^1 \left(D^3\ell(w_z) + \epsilon\left(D^3\ell(w_z + x\epsilon e_1) - D^3\ell(w_z)\right)\right)\,dx\,(e_1,\cdot,\cdot) \\
&= D^2\ell(w_z)(\cdot,\cdot) + D^3\ell(w_z)(\epsilon e_1,\cdot,\cdot) + \epsilon\int_0^1 \left(D^3\ell(w_z + x\epsilon e_1) - D^3\ell(w_z)\right)\,dx\,(e_1,\cdot,\cdot) \\
&= D^2\ell(w_z)(\cdot,\cdot) + D^3\ell(w_z)(\epsilon e_1,\cdot,\cdot) + \frac{\epsilon^2 B}{2}E(\cdot,\cdot),
\end{aligned}$$

where the linear operator E satisfies $\|E\| \leq 1$. Hence (using E to also denote the corresponding matrix),

$$\begin{aligned}
\nabla^2 \ell(w_z + \epsilon e_1) &= \nabla^2 \ell(w_z) + \sum_{i,j} D^3 \ell(w_z)(\epsilon e_1, e_i, e_j) e_i \otimes e_j + \frac{\epsilon^2 B}{2} E \\
&= \sum_i \lambda_i e_i \otimes e_i + D^3 \ell(w_z)(\epsilon e_1, e_1, e_1) e_1 \otimes e_1 \\
&\quad + \sum_{i>1} D^3 \ell(w_z)(\epsilon e_1, e_1, e_i)(e_1 \otimes e_i + e_i \otimes e_1) \\
&\quad + \sum_{i>1, j>1} D^3 \ell(w_z)(\epsilon e_1, e_i, e_j) e_i \otimes e_j + \frac{\epsilon^2 B}{2} E.
\end{aligned}$$

Integrating from $x = 0$ to $x = \epsilon$, we have

$$\begin{aligned}
&\nabla \ell(w_z + \epsilon e_1) \\
&= \nabla \ell(w_z) + \int_0^\epsilon \nabla^2 \ell(w_z + x e_1) e_1 \, dx \\
&= \int_0^\epsilon \left(\sum_i \lambda_i e_i \otimes e_i + D^3 \ell(w_z)(x e_1, e_1, e_1) e_1 \otimes e_1 \right. \\
&\qquad + \sum_{i>1} D^3 \ell(w_z)(x e_1, e_1, e_i)(e_1 \otimes e_i + e_i \otimes e_1) \\
&\qquad \left. + \sum_{i>1, j>1} D^3 \ell(w_z)(x e_1, e_i, e_j) e_i \otimes e_j + \frac{x^2 B}{2} E \right) e_1 \, dx \\
&= \int_0^\epsilon \left(\lambda_1 e_1 + D^3 \ell(w_z)(x e_1, e_1, e_1) e_1 + \sum_{i>1} D^3 \ell(w_z)(x e_1, e_1, e_i) e_i + \frac{x^2 B}{2} E e_1 \right) dx \\
&= \epsilon \lambda_1 e_1 + \frac{\epsilon^2}{2} \sum_i D^3 \ell(w_z)(e_1, e_1, e_i) e_i + \frac{\epsilon^3 B}{6} E e_1.
\end{aligned}$$

Substituting $\epsilon = s_t(\beta_1/\lambda_1 + \rho)$, the first term is

$$\begin{aligned}
\epsilon \lambda_1 e_1 &= s_t \left(\frac{\beta_1}{\lambda_1} + \rho \right) \lambda_1 e_1 \\
&= s_t \left(\frac{\eta \rho \lambda_1^2}{2 - \eta \lambda_1} + \rho \lambda_1 \right) e_1 \\
&= \frac{2 \rho \lambda_1 s_t}{2 - \eta \lambda_1} e_1 \\
&= \frac{2 \beta_1 s_t}{\eta \lambda_1} e_1.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\eta\nabla\ell(\tilde{w}_u)\\
&= \frac{2\beta_1 s_t}{\lambda_1}e_1 + \eta\frac{(\beta_1/\lambda_1+\rho)^2}{2}\sum_i D^3\ell(w_z)(e_1,e_1,e_i)e_i + \eta s_t\frac{(\beta_1/\lambda_1+\rho)^3B}{6}Ee_1\\
&= \frac{2\beta_1 s_t}{\lambda_1}e_1 + \frac{\eta(\beta_1/\lambda_1+\rho)^2}{2}\nabla\lambda_{\max}(\nabla^2\ell(w_z)) + \frac{\eta(\beta_1/\lambda_1+\rho)^3B}{6}\zeta, \qquad (16)
\end{aligned}$$

where $\|\zeta\| \le 1$.

Now, we turn to bounding $||\nabla\ell(\tilde{w}_u) - \nabla\ell(w_u)||$. (We will show that $\tilde{w}_u$ and w_u are both close to w_z, so that the operator norm of the Hessian is not too big between them, and then we will show that they are close to one another.) First, by the triangle inequality,

$$\max\{||\tilde{w}_u - w_z||, ||w_u - w_z||\} \le \beta_1/\lambda_1 + \rho.$$

Since $D^3\ell$ is B-Lipschitz, this implies that, for every w on the path from w_u to $\tilde{w}_u$,

$$||\nabla^2\ell(w)|| \le \lambda_1 + B(\beta_1/\lambda_1+\rho). \qquad (17)$$

Furthermore, we have

$$||w_u - \tilde{w}_u|| = \rho\left\|s_t e_1 - \frac{\nabla\ell(w_t)}{||\nabla\ell(w_t)||}\right\|. \qquad (18)$$

Next,

$$\begin{aligned}
\nabla\ell(w_t) &= \nabla\ell\left(w_z + \frac{s_t\beta_1}{\lambda_1}e_1\right)\\
&= \nabla\ell(w_z) + \int_0^1 \nabla^2\ell\left(w_z + x\left(\frac{s_t\beta_1}{\lambda_1}e_1\right)\right)\left(\frac{s_t\beta_1}{\lambda_1}e_1\right)dx\\
&= \nabla^2\ell(w_z)\left(\frac{s_t\beta_1}{\lambda_1}e_1\right)\\
&\qquad + \int_0^1\left(\nabla^2\ell\left(w_z + x\left(\frac{s_t\beta_1}{\lambda_1}e_1\right)\right) - \nabla^2\ell(w_z)\right)\left(\frac{s_t\beta_1}{\lambda_1}e_1\right)dx\\
&= s_t\beta_1 e_1 + \frac{B\beta_1^2}{2\lambda_1^2}\xi
\end{aligned}$$

for $\xi \in \mathbb{R}^d$ with $\|\xi\| \le 1$.

This implies $||\nabla\ell(w_t)|| \ge \beta_1 - \frac{B\beta_1^2}{2\lambda_1^2}$, which in turn implies

$$\begin{aligned}
\left\|\frac{\nabla\ell(w_t)}{\|\nabla\ell(w_t)\|} - s_t e_1\right\| &= \left\|\left(\frac{\beta_1}{\|\nabla\ell(w_t)\|} - 1\right)s_t e_1 + \frac{B\beta_1^2}{2\lambda_1^2\|\nabla\ell(w_t)\|}\xi\right\|\\
&\le \frac{B\beta_1/(2\lambda_1^2)}{1 - B\beta_1/(2\lambda_1^2)} + \frac{B\beta_1}{2\lambda_1^2(1 - B\beta_1/(2\lambda_1^2))}\\
&= \frac{2B\beta_1}{2\lambda_1^2 - B\beta_1}.
\end{aligned}$$

Recalling (18),

$$||w_u - \tilde{w}_u|| \le \frac{2B\beta_1\rho}{2\lambda_1^2 - B\beta_1},$$

and by (17), this implies

$$||\nabla\ell(w_u) - \nabla\ell(\tilde{w}_u)|| \le \frac{2B\beta_1\rho\,(\lambda_1 + B\beta_1/\lambda_1 + B\rho)}{2\lambda_1^2 - B\beta_1}.$$

Putting this together with (16) and (15), there is a ζ with $||\zeta|| \le 1$ for which

$$\begin{aligned} w_{t+1} - w_t = &- \frac{2\beta_1 s_t}{\lambda_1} e_1 - \frac{\eta(\beta_1/\lambda_1 + \rho)^2}{2} \nabla\,\lambda_{\max}(\nabla^2\ell(w_z)) \\ &+ \left(\frac{\eta(\beta_1/\lambda_1 + \rho)^3 B}{6} + \frac{2B\eta\beta_1\rho\,(\lambda_1 + B\beta_1/\lambda_1 + B\rho)}{2\lambda_1^2 - B\beta_1} \right) \zeta \end{aligned}$$

which, substituting the value of β_1 and applying $B\eta\rho \le 1$ and $\eta\lambda_1 < 1$, implies

$$\begin{aligned} w_{t+1} - w_t = &-2\frac{\eta\rho\lambda_1 s_t}{2 - \eta\lambda_1} e_1 - \frac{\eta}{2}\left(\frac{\eta\rho\lambda_1}{2 - \eta\lambda_1} + \rho\right)^2 \nabla\,\lambda_{\max}(\nabla^2\ell(w_z)) \\ &+ \eta\rho^2 \left(\frac{(1 + \eta\lambda_1)^3\rho}{6} + 2(2\lambda_1 + B\rho)\eta \right) B\zeta. \end{aligned}$$

■

6. Additional Simulations

Figure 3 compares the trajectories of SAM (in blue) and batch gradient descent (in green) applied to $\frac{w_1^2}{1+w_2^2/2} + w_2^2/2$. It may be helpful to think of this objective as a perturbation of the quadratic objective $w_1^2 + w_2^2/2$, that has the same minimum, but, as w_2 moves away from zero, is less sharp, in the sense that its Hessian has a smaller operator norm. When SAM and GD are both started $(0.1, 0.1)$, with $\eta = 1/5$ and $\rho = 1$, GD dives toward the minimum of 0, where SAM's oscillation drives it toward less sharp solutions with larger objective values.

Figure 4 compares the trajectories of SAM and GD in the same setting, except from the initial solution $(1, 1)$. SAM behaves similarly to GD until they get close to the origin, where SAM's oscillations carry it to a less sharp minimum with a larger objective value.

Figure 5 compares the trajectories of SAM and SGD, where each stochastic gradient is obtained by perturbing the gradient by a sample from $\mathcal{N}(0, \sigma^2 I)$, for $\sigma = \rho/(2 - \eta)$. The perturbed gradients make the iterates of SGD sample a mix of solutions with varying smoothness, where SAM systematically drifts toward less sharp solutions.

7. Conclusions and Open Problems

Our main result, Theorem 1, shows that SAM with a convex quadratic objective converges to a cycle that bounces across the minimum in the direction with the largest curvature.

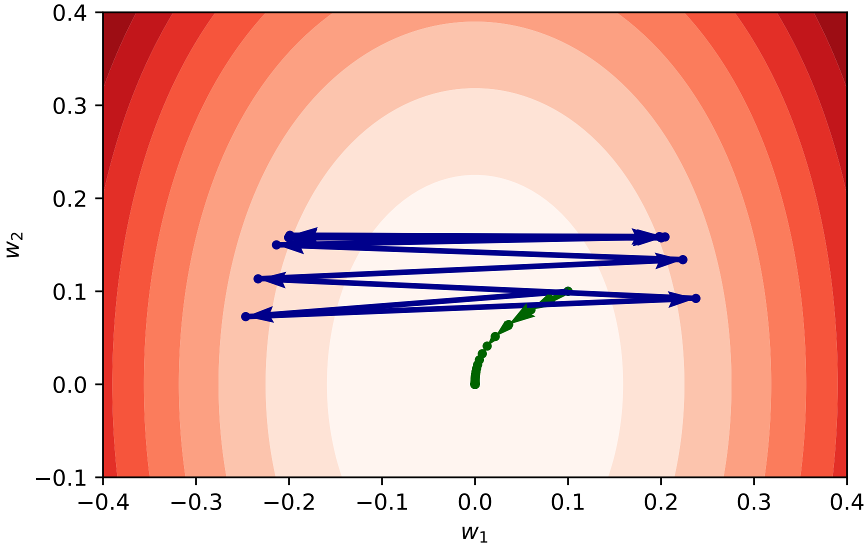


Figure 3: SAM (in blue) and gradient descent (in green) applied to $\frac{w_1^2}{1+w_2^2/2} + w_2^2/2$ from an initial solution of $(0.1, 0.1)$ with $\eta = 1/5$ and $\rho = 1$.

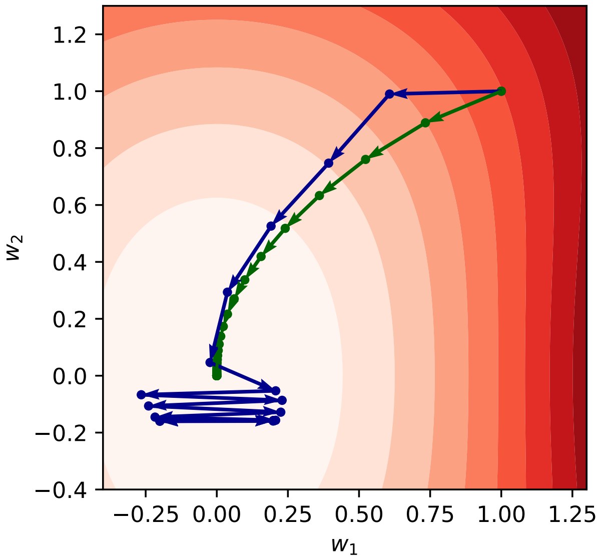


Figure 4: SAM (in blue) and gradient descent (in green) applied to $\frac{w_1^2}{1+w_2^2/2} + w_2^2/2$ from an initial solution of $(1, 1)$ with $\eta = 1/5$ and $\rho = 1$.

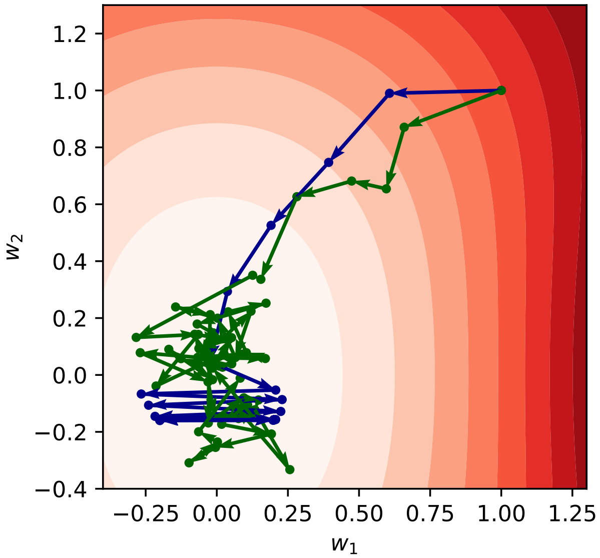


Figure 5: SAM (in blue) and SGD (in green) applied to $\frac{w_1^2}{1+w_2^2/2} + w_2^2/2$ from an initial solution of $(1, 1)$ with $\eta = 1/5$, $\rho = 1$ and $\sigma = \rho/(2 - \eta)$.

Theorem 21 shows that for a locally quadratic loss, these oscillations allow gradient descent on the spectral norm of the Hessian of the loss. SAM uses one additional gradient measurement per iteration to compute a specific third derivative: the gradient of the second derivative in the leading eigenvector direction.

Without the assumption that $\lambda_1 > \lambda_2$, Theorem 1 would necessarily be more complex, since, informally, if $\lambda_1 = \lambda_2$, all solutions in the span of e_1 and e_2 are equivalent. It should not be hard to remove this assumption while complicating some of the proofs, but without significant changes to the main ideas.

This work raises several natural questions. First, how is the generalization behavior affected by drifting towards wide minima? There have been several empirical studies of stochastic gradient methods for deep networks that suggest favorable generalization performance of wide minima (Keskar et al., 2016; Chaudhari et al., 2019). There have been some analyses aimed at understanding this phenomenon based on information theoretic arguments (Hinton and van Camp, 1993; Hochreiter and Schmidhuber, 1997; Negrea et al., 2019) and PAC-Bayes arguments (Langford and Caruana, 2001; Dziugaite and Roy, 2017). It is clear that any argument about generalization properties must take account of how an algorithm solves an optimization problem over a parameterized class of functions, since wide minima are a property of a parameterization (Dinh et al., 2017).

Second, how does gradient descent on the spectral norm of the Hessian behave, particularly in the highly overparameterized setting of deep networks? When other optimization tools, such as momentum, are incorporated, how does this affect the behavior of SAM? What is the nature of SAM's solutions for losses, like the logistic loss, that are minimized at infinity?

On the technical side, it is straightforward to extend Lemma 10 to a local version, showing that SAM with a locally quadratic loss converges to a neighborhood of the stationary points of a function J defined in terms of the Hessian. It is less straightforward to show that SAM avoids the suboptimal stationary points of J. It seems likely that this is true for a stochastic version of the SAM updates, and the techniques developed by Ge et al. (2015); Fang et al. (2019) should be useful here, which could lead to a nonasymptotic counterpart of results of Wen et al. (2023) for a stochastic (batch-size 1) version of SAM.

Finally, can other higher derivatives be computed in the same parsimonious way as SAM? Are there related minimization methods that target other kinds of minima, for instance, by optimizing other measures of width of a minimum?

Acknowledgments

Thanks to Hossein Mobahi for helpful comments. PB gratefully acknowledges the support of the NSF through grants DMS-2023505 and DMS-2031883 and of Simons Foundation award #814639.

References

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *ICML*, pages 247–257, 2022.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024, 2022.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477, 2021.

Dara Bahri, Hossein Mobahi, and Yi Tay. Sharpness-aware minimization improves language model generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 7360–7371, 2022.

David Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2020.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

Gaspard Beugnot, Julien Mairal, and Alessandro Rudi. On the benefits of large learning rates for kernel methods. In *Conference on Learning Theory*, pages 254–282, 2022.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2020.

Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*, 2022.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2022.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*, 2017.

Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex SGD escaping from saddle points. In *Conference on Learning Theory*, pages 1192–1234, 2019.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Conference on Computational Learning Theory*, page 5–13, 1993.

Sepp Hochreiter and Jürgen Schmidhuber. Flat Minima. *Neural Computation*, 9(1):1–42, 1997.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2016.

John Langford and Rich Caruana. (Not) bounding the true error. In *Advances in Neural Information Processing Systems*, 2001.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Advances in Neural Information Processing Systems*, 2019.

Samuel L Smith, Benoit Dherin, David Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How sharpness-aware minimization minimizes sharpness? In *The Eleventh International Conference on Learning Representations*, 2023.