Flat Minima and Generalization: Insights from Stochastic Convex Optimization

Matan Schliserman[*] SCHLISERMAN@MAIL.TAU.AC.IL
Shira Vansover-Hager[*] SHIRAV@MAIL.TAU.AC.IL
Tel Aviv University

Tomer Koren TKOREN@TAUEX.TAU.AC.IL
Tel Aviv University and Google Research

Abstract

Understanding the generalization behavior of learning algorithms is a central goal of learning theory. A recently emerging explanation is that learning algorithms are successful in practice because they converge to flat minima, which have been consistently associated with improved generalization performance. In this work, we study the link between flat minima and generalization in the canonical setting of stochastic convex optimization with a non-negative, β-smooth objective. Our first finding is that, even in this fundamental and well-studied setting, flat empirical minima may incur trivial $\Omega(1)$ population risk while sharp minima generalizes optimally. Then, we show that this poor generalization behavior extends to two natural ”sharpness-aware” algorithms originally proposed by Foret et al. [25], designed to bias optimization toward flat solutions: Sharpness-Aware Gradient Descent (SA-GD) and Sharpness-Aware Minimization (SAM). For SA-GD, which performs gradient steps on the maximal loss in a predefined neighborhood, we prove that while it successfully converges to a flat minimum at a fast rate, the population risk of the solution can still be as large as $\Omega(1)$, indicating that even flat minima found algorithmically using a sharpness-aware gradient method might generalize poorly. For SAM, a computationally efficient approximation of SA-GD based on normalized ascent steps, we show that although it minimizes the empirical loss, it may converge to a sharp minimum and also incur population risk $\Omega(1)$. Finally, we establish population risk upper bounds for both SA-GD and SAM using algorithmic stability techniques.

1. Introduction

Understanding the generalization behavior of modern learning algorithms has become a central focus of theoretical machine learning. This interest is motivated by the observation that in heavily overparameterized deep neural networks, the training objective admits many global optima that perfectly fit the data [73]; yet, while some of these minimizers generalize poorly, others—typically those to which common optimization algorithms converge—generalize well [49, 50, 73]. These observations naturally raise the fundamental question of what theoretical and algorithmic conditions ensure that minimizers generalize well.

One prominent condition that has received significant attention is the *flatness* of the minimum. Flat minima—those that remain (approximate) minimizers under small parameter perturbations—have been consistently associated with better generalization, while sharper, non-flat minima are linked with worse out-of-sample performance [21, 31, 33, 58]. This insight has motivated a va-

[*] Equal contribution.

riety of methods that encourage solutions in flat regions of the loss landscape, rather than sharp ones [4, 19, 20, 25, 30, 35, 37, 40, 41, 43, 61, 69, 72, 74–76]. In particular, Foret et al. [25] introduced the Sharpness-Aware Minimization (SAM) approach, which reformulates the standard optimization problem as minimizing the *Sharpness-Aware Empirical Risk (SAER)*, defined as $F_S^r(w) = \max_{\|v\| \leq r} F_S(w+v)$ where F_S is the empirical risk over a sample S and r is a *perturbation radius* parameter. This approach encourages solutions robust to parameter perturbations, thus corresponding to flatter minima.

Despite the success of SAM, as well as of other sharpness-aware methods [7, 13, 25, 32, 38], the theoretical link between flatness and generalization remains not fully understood. While some works show that in certain non-convex regimes the flatness of an arbitrary minimizer does not affect generalization [e.g., 18, 68], it is unclear whether this also holds for concrete optimization methods that explicitly aim to find flat minima. For such methods, existing analyses either provide only empirical evidence [5, 53, 68], establish problem parameters-dependent generalization bounds [25, 50, 52, 64, 65], or restrict attention to quadratic or strongly convex objectives [14, 62]. As a result, it remains unclear whether and under which conditions finding a flat empirical minimum using such algorithms does in fact lead to improved generalization, or how the generalization guarantees of these practical methods compare to those of standard optimization algorithms such as gradient descent (GD) and stochastic gradient descent (SGD).

In this paper, we aim to gain insight into the relationship between flatness and generalization by studying the above questions within the framework of Stochastic Convex Optimization (SCO): a fundamental and extensively studied theoretical model widely used to analyze stochastic optimization algorithms. SCO is particularly well-suited for such a study, as it is well-known that SCO problems can admit multiple empirical minimizers, not all of which are guaranteed to generalize well [24, 56]. We focus on the regime where the loss functions are non-negative and β-smooth;[1] in this setting, gradient methods such as GD and SGD are known to generalize optimally [27, 51], as opposed to the wider convex non-smooth setting [2, 44, 55, 63]. Within this SCO framework, we impose the additional assumption that f admits at least one flat minimum, i.e., a minimizer such that the loss remains constant within a ball of radius ρ around it (we call such a minimizer a *ρ-flat minimum*). To capture this formally, we introduce a strong flatness condition (see Definition 1), and analyze the generalization performance of several natural algorithms under this condition.

Our contributions shed light on the extent to which flatness relates to generalization in SCO. We construct examples showing that flat empirical minima can generalize poorly, demonstrating that minimizing the Sharpness-Aware Empirical Risk (SAER) does not in itself guarantee good generalization. First, we present an SCO instance in which a flat empirical risk minimizer (ERM) generalizes poorly, while within the same setting, a sharp ERM generalizes well. Then, we show that this poor generalization behavior extends to two natural "sharpness-aware" algorithms originally proposed by Foret et al. [25], designed to bias optimization toward flat solutions: Sharpness-Aware Gradient Descent (SA-GD)[2] and Sharpness-Aware Minimization (SAM). For SA-GD, we prove that it indeed converges to a flat minimum, however, there are instances where it converges to solutions that generalize strictly worse compared to those found by standard GD and SGD, which are known to generalize optimally in the same setting. These results indicate that even flat minima found algorithmically using a sharpness-aware gradient method might generalize poorly. For SAM we observe a sharper contrast: although it minimizes the empirical risk, it does not necessarily minimize

[1] A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if $\|\nabla f(v) - \nabla f(u)\|_2 \leq \beta \|v - u\|_2$ for all $u, v \in \mathbb{R}^d$.

[2] This algorithm was introduced in [25] without being explicitly named; we refer to it as SA-GD here for conciseness.

sharpness as it may converge to a non-flat minimum, and similarly to SA-GD, we show it might converge to minima with poor generalization compared to (S)GD. These results provide insight into possible limitations of sharpness-aware approaches in terms of the flatness of the solution found and its out-of-sample performance relative to (S)GD.

Summary of contributions. In more detail, we make the following technical contributions. (The bounds presented below describe the dependence on the number of iterations T, the number of training examples n, the step size η, the smoothness parameter β, the flatness radius of the loss minimizer ρ, and the perturbation size r.)

(i) We introduce a strong flatness condition assuming the existence of a perfectly flat minimum of radius ρ (Definition 1). For Sharpness-Aware ERM (SA-ERM), even under this strong condition, we construct a smooth SCO problem where the empirical risk admits a flat minimizer with population risk $\Omega(1)$, while a non-flat minimizer achieves optimal generalization (Theorem 1).

(ii) For the SA-GD algorithm [25], we prove an empirical optimization bound $O(1/\eta T + \max(r - \rho, 0)^2)$, implying that with $\eta \simeq 1/\beta$ and $r \simeq \rho$, SA-GD converges to a $\Theta(\rho)$-flat minimum at rate $O(1/T)$. In contrast, we establish a lower bound of $\Omega(\eta^2(r-\rho)^2 T)$ on the population loss of SA-GD for $r \gtrsim \rho$, showing that SA-GD may generalize poorly even when converging to flat minima. In particular, tuning the algorithm with $\eta \simeq 1/\beta$ and $r \gtrsim \rho + 1/\sqrt{T}$ can lead to a population risk of $\Omega(1)$ (Theorems 2 and 4).

(iii) For SAM [25], we obtain the same bound $O(1/(\eta T) + \max(r - \rho, 0)^2)$ for the empirical risk, but also show a convex, smooth case where SAM converges to a sharp minimum, i.e., it fails to minimize the SAER. As for generalization, we establish a population lower bound of $\Omega(\eta^2 r^2 T)$ in the case $\rho = 0$, which implies a trivial risk of $\Omega(1)$ when $\eta \simeq 1/\beta$ and $r \gtrsim 1/\sqrt{T}$, or when $\eta \simeq 1/\sqrt{T}$ and $r = \Theta(1)$, regimes where SAM minimizes the empirical risk (Theorems 3, 5 and 6).

(iv) Finally, using algorithmic stability, we prove population upper bounds for SA-GD and SAM under ρ-flatness. In particular, for $T = n$, $\eta \simeq 1/\beta$, and $r \lesssim \rho + 1/\sqrt{T}$, the bounds are of order $O(1/n + r^2 n)$. For the precise statements, see Theorems 7 and 8 in Appendix C.

To our knowledge, these results are the first to formally address the connection between flatness and generalization in the convex regime, and they bear some interesting implications. On the positive side, they provide the first indication that sharpness-aware methods converge at a dimension-independent fast $O(1/T)$ rate in terms of empirical risk for general convex optimization, despite the SAER objective being non-smooth, and this convergence can further benefit from flatness of the objective. On the negative side, our results show that even in the basic convex and smooth regime, a sharp empirical minimum may generalize better than a flat one, and this can occur when the flat empirical risk minimizer is selected arbitrarily, e.g., by the SA-ERM algorithm, or algorithmically, by the SA-GD algorithm. Furthermore, our findings highlight that optimization methods explicitly designed to locate flat minima, such as SA-GD and SAM, may converge to solutions that generalize poorly. In contrast, standard gradient-based methods like GD and SGD are known to achieve optimal generalization in this setting when using the optimization-optimal step size $\eta \simeq 1/\beta$ [39, 51].

2. Problem setup

We study the generalization properties of flat minima in the framework of (smooth) *Stochastic Convex Optimization* (SCO). In this setting, there exists a population distribution $\mathcal{D}$ over an instance space $\mathcal{Z}$, and a loss function $f : W \times \mathcal{Z} \to \mathbb{R}$ defined on a convex domain $W \subseteq \mathbb{R}^d$. For any fixed instance $z \in \mathcal{Z}$, the function $f(\cdot, z)$ is assumed to be non-negative, convex, and β-smooth ($\beta > 0$) with respect to its first argument w. The learning goal is to minimize the *population risk*, defined as the expected loss over $\mathcal{D}$,

$$F(w) \coloneqq \mathbb{E}_{z\sim\mathcal{D}}[f(w, z)]. \tag{1}$$

Since $\mathcal{D}$ is unknown, learning algorithms instead use a finite i.i.d. sample $S = \{z_1, \ldots, z_n\}$ drawn from $\mathcal{D}$. A common approach is to minimize the *empirical risk* over S, given by

$$F_S(w) \coloneqq \tfrac{1}{n} \textstyle\sum_{i=1}^n f(w, z_i). \tag{2}$$

A main focus of this paper is on objective functions that admit *flat minima*, formalized as follows.

Definition 1 (ρ-flatness) We say that $w^\star \in \mathbb{R}^d$ is a *ρ-flat minimum* (for $\rho \geq 0$) of a non-negative function $f : \mathbb{R}^d \to \mathbb{R}$ if for every $w \in \mathbb{R}^d$ with $\|w - w^\star\| \leq \rho$, it holds that $f(w) = 0$. If such a ρ-flat minimum exists for f, we also say that f is ρ-flat; the maximal ρ satisfying this condition is called the *flatness radius* of f.

Note that this is a rather strong notion of flatness: it in particular implies that the empirical minimization problem with a ρ-flat F_S is *realizable* (i.e., there exists $w^\star$ such that $f(w^\star, z_i) = 0$ for almost all $z_i \in S$) and further that F_S is *perfectly flat* in a neighborhood of $w^\star$. Since our goal is to understand the relationship between flatness and generalization, we find it more informative to analyze this connection under the most stringent and unambiguous condition of flatness. In particular, imposing such a condition makes any negative results (i.e., lower bounds) only *stronger*, since they hold even under the most favorable notion of flatness.

With the above notion of flatness in mind, we focus on three natural algorithms:

- ***Sharpness-Aware Empirical Risk Minimization (SA-ERM).*** The first (meta-)algorithm is a natural, "Sharpness-Aware" variant of ERM that computes, given a parameter $r > 0$:

  $$w_S \in \operatorname*{arg\,min}_{w\in W} F_S^r(w), \quad \text{where} \quad F_S^r(w) = \max_{v:\ \|v\|\leq r} F_S(w + v). \tag{3}$$

  Namely, it outputs a minimizer of the *sharpness-aware empirical risk* (SAER) with radius r, which we denote by F_S^r. The idea here is that, if the empirical risk F_S is ρ-flat and $r \leq \rho$, then *any* minimizer of the SAER is also a r-flat minimum of the original empirical risk F_S.

- ***Sharpness-Aware Gradient Descent (SA-GD).*** The second algorithm is a first-order instantiation of SA-ERM, proposed in [25], obtained by running gradient descent on the SAER objective. Starting from $w_1 \in W$ and given parameters $\eta, r > 0$, it takes steps for $t = 1, \ldots, T$ of the form:

  $$w_{t+1} = w_t - \eta \nabla F_S(w_t + v_t), \quad \text{where} \quad v_t \in \operatorname*{arg\,max}_{v:\ \|v\|\leq r} F_S(w_t + v). \tag{4}$$

- ***Sharpness-Aware Minimization (SAM).*** The third algorithm is the original SAM algorithm proposed in [25] as a computationally efficient approximation of SA-GD. SAM circumvents the explicit maximization over v in Eq. (4) by replacing v_t with the normalized gradient at w_t. Thus, starting from $w_1 \in W$ and given $\eta, r > 0$, the updates of SAM for $t = 1, \ldots, T$ take the form

  $$w_{t+1} = w_t - \eta \nabla F_S\left(w_t + r\frac{\nabla F_S(w_t)}{\|\nabla F_S(w_t)\|}\right). \tag{5}$$

3. Overview of results and techniques

In this section, we give an overview of our generalization lower bounds for the algorithms we consider (SA-ERM, SA-GD, SAM). Due to space limitations, the formal statements of the optimization and generalization upper bounds for SA-GD and SAM are deferred to Appendices B and C.

Generalization of SA-ERM. Our first result provides a lower bound on the generalization of SA-ERM, showing that an arbitrary minimizer of SAER may exhibit trivial $\Omega(1)$ population risk. This result highlights not only the limitations of the SA-ERM algorithm in general smooth SCO but also the effect of the loss landscape on generalization. Even when the loss is convex and β-smooth and under the arguably strongest notion of flatness (Definition 1), a flat minimum of the empirical risk may generalize poorly, while a sharp minimum of the same function generalizes optimally.

Theorem 1 *For every $n \in \mathbb{N}$ and $0 \le \rho \le \frac{1}{2}$, let $d = 2^n + 1$ and define $W = \{x \in \mathbb{R}^d : \|x\| \le 1\}$. Then there exist an instance set $\mathcal{Z}$, a distribution $\mathcal{D}$ over $\mathcal{Z}$, and a loss function $f : W \times \mathcal{Z} \to \mathbb{R}$ that is convex, 1-Lipschitz, 1-smooth and ρ-flat, such that with probability at least $\frac{1}{2}$ over the training set S, there exist $w^{(1)}, w^{(2)} \in \arg\min_{w \in W} F_S(w)$ satisfying:*

- *(i) for every $r \ge 0$, it holds that $w^{(1)} \in \arg\min_{w \in W} F_S^r(w)$. In particular, if $r \le \rho$ then $w^{(1)}$ is an r-flat minimum of F_S;*
- *(ii) $w^{(2)}$ is a sharp minimum, in the sense that $F_S^\delta(w^{(2)}) \ge F_S(w^{(2)}) + \frac{1}{2}\delta^2$ for all $\delta > 0$.*[3]
- *(iii) we have $F(w^{(1)}) - F(w^\star) = \Omega(1)$, while $F(w^{(2)}) - F(w^\star) = 0$.*

Our construction of the smooth SCO instance builds on previous results in SCO, which show that arbitrary minimizers of the empirical risk may overfit [24, 56]. The main technical challenge is that in those prior constructions the ERM exhibiting poor generalization is not a flat minimizer. Thus, our key challenge in the proof is to transform this ERM into a flat minimizer. A simple observation is that the function $h : \mathbb{R} \to \mathbb{R}$, $h(x) = \frac{1}{2}\max(x - \rho, 0)^2$ is 1-smooth and ρ-flat for $\rho \le \frac{1}{2}$. Given this observation, we construct the instance where SA-ERM generalizes poorly by composing this function with a careful variant of the construction from Shalev-Shwartz et al. [56].

Generalization of SA-GD. Our next result is a lower bound on the population risk of SA-GD. From Theorem 4, we know that in the ρ-flat regime, when $r - \rho \le \frac{1}{\sqrt{T}}$, SA-GD minimizes the SAER and converges to a flat minimum. Given Theorem 1, which shows that an arbitrary flat minimum may fail to generalize, a natural question is whether a flat minimum attained by a practical algorithm such as SA-GD be guaranteed to generalize well.

In the following we show that this does not necessarily hold. Even when SA-GD converges to a flat minimum, for instance, when $\eta \approx \frac{1}{\beta}$ and $r \approx \rho + \frac{1}{\sqrt{T}}$, its population risk can still be $\Omega(1)$.

Theorem 2 *For every $n, T \in \mathbb{N}, , \eta > 0, r \ge 0, \rho < r\big(1 - \frac{3}{3+\eta\sqrt{T}}\big)$, assume $\eta(r - \rho) \le \frac{1}{\sqrt{T}}$, let $d = 2^n T$ and define $W = \{x \in \mathbb{R}^d : \|x\| \le 1\}$. Then there exists an instance set $\mathcal{Z}$, a distribution $\mathcal{D}$ over $\mathcal{Z}$, function $f : W \times \mathcal{Z} \to \mathbb{R}$ that is convex 1-smooth, 1-Lipschitz and ρ-flat, such that for a training set S it holds that with probability at least $\frac{1}{2}$, running SA-GD for T steps yields for every $\tau \in [T]$ suffix average $\widehat{w}_\tau = \frac{1}{T-\tau+1}\sum_{t=\tau}^{T} w_t$:*

$$F(\widehat{w}_\tau) - F(w^\star) = \Omega(\eta^2 (r - \rho)^2 T).$$

[3] This condition means that in every neighborhood of the minimizer there exists a point with large F_S. The inequality is the tightest possible: due to 1-smoothness, any minimizer $w^\star$ of F_S satisfies $F_S^\delta(w^\star) \le F_S(w^\star) + \frac{1}{2}\delta^2$ for all $\delta > 0$.

In particular, if $\eta = \Theta(1)$, $\rho \le \frac{\eta}{3}$ and $r = \rho + \frac{1}{\sqrt{T}}$ this yields a lower bound of $\Omega(1)$.

The main technical challenge in the proof is that, in the non-smooth setting, prior constructions (e.g., [2, 36, 44, 55, 63]) exploit non-smoothness to shape the algorithm's dynamics, whereas in the smooth setting such an approach is not possible. Instead, our key idea is to control the sequence of maximizers $\{v_t \in \arg\max_{\|v\| \le r} F_S(w_t + v)\}_{t=1}^T$ to direct the dynamics toward a spurious ERM. For this, we base our hard instance on the construction for SA-ERM given in Theorem 1. In that construction, in the first iteration we have $v_1 = re_i$, where e_i corresponds to the spurious ERM from Theorem 1. As a result, SA-GD makes a single step of size ηr toward this bad ERM. The remaining challenge is to ensure that the algorithm takes T such steps in this direction. To achieve this, we construct a new loss function that applies the loss from Theorem 1 in T orthogonal subspaces. In this way, since v_t is chosen in a different subspace at each iteration t, the algorithm makes a single step in each subspace and eventually converges to a bad ERM.

Generalization of SAM. Finally, we turn to discuss the generalization guarantees of SAM, a well-studied and practically relevant algorithm introduced by [25] as a computationally efficient approximation of SA-GD. In the following lower bound, we show that SAM can exhibit poor generalization in SCO under the realizable setting ($\rho = 0$), leaving the ρ-flat case ($\rho \gg 0$) for future work.

Theorem 3 *Given $n \ge 6, T \ge 6, \eta, r > 0$ such that $\eta r \le 1/2\sqrt{T}$, let $d = 2^n T$ and $W = \{w \in \mathbb{R}^d : \|w\| \le 1\}$. Then there exists an instance set $\mathcal{Z}$, a distribution $\mathcal{D}$ over $\mathcal{Z}$, a convex 6-smooth 7-Lipschitz and realizable function $f : W \times \mathcal{Z} \to \mathbb{R}$ such that for a training set S with probability at least $\frac{1}{3}$ running SAM for T steps with trajectory $\{w_t\}_{t=1}^T$, yields for every $\tau \in [T]$ suffix average $\widehat{w}_\tau = \frac{1}{T-\tau+1}\sum_{t=\tau}^T w_t$:*

$$F(\widehat{w}_\tau) - F(w^\star) = \Omega(\eta^2 r^2 T).$$

For the proof, as in the construction for SA-GD in Theorem 2, we need to use the perturbations (caused by the normalized ascent steps) of the algorithm to direct it toward a spurious ERM. The main difficulty in this context is that, in the previous construction, the algorithm is initialized at $w_1 = 0$, which is already a minimizer of the empirical risk. As a result, if we were to apply the same approach, SAM would remain at initialization throughout training and thus generalize well.

To overcome this challenge, our key idea is to exploit the normalization of the ascent step, which can amplify small perturbations into significant effects. In particular, our construction begins with a sufficiently small quadratic function in the first orthogonal subspace. Although the gradients of this function at the initialization point are small, the normalization step amplifies them, producing a progress of ηr toward the bad ERM in this subspace. To achieve T such steps, we carefully design an additional mechanism with sufficiently small loss in each orthogonal subspace, such that for every pair of consecutive subspaces, the previous step induces a following step in the next orthogonal subspace. This allows the algorithm to make progress of ηr in T different subspaces and converge to an ERM that generalizes poorly.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101078075). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work received additional support from the Israel Science Foundation (ISF, grant numbers 2549/19 and 3174/23), a grant from the Tel Aviv University Center for AI and Data Science (TAD) and from the Len Blavatnik and the Blavatnik Family foundation.

In addition, this work was partially supported by the TAD Excellence Program for Doctoral Students in Artificial Intelligence and Data Science from the Tel Aviv University Center for AI and Data Science (TAD).

References

[1] Atish Agarwala and Yann Dauphin. SAM operates far from home: eigenvalue regularization as a dynamical phenomenon. In *ICML*, 2023.

[2] Idan Amir, Tomer Koren, and Roi Livni. SGD generalizes better than gd (and regularization doesn't help). In *Conference on Learning Theory*, pages 63–92. PMLR, 2021.

[3] Maksym Andriushchenko and Nicolas Flammarion. Towards Understanding Sharpness-Aware Minimization. In *ICML*, 2022.

[4] Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. In *NeurIPS*, 2023.

[5] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. *arXiv preprint arXiv:2302.07011*, 2023.

[6] Amit Attia, Matan Schliserman, Uri Sherman, and Tomer Koren. Fast last-iterate convergence of sgd in the smooth interpolation regime. *arXiv preprint arXiv:2507.11274*, 2025.

[7] Dara Bahri, Hossein Mobahi, and Yi Tay. Sharpness-aware minimization improves language model generalization. *arXiv preprint arXiv:2110.08529*, 2021.

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

[9] Kayhan Behdin and Rahul Mazumder. Sharpness-aware minimization: An implicit regularization perspective. *arXiv preprint arXiv:2302.11836*, 2023.

[10] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

[11] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[12] Daniel Carmon, Roi Livni, and Amir Yehudayoff. The sample complexity of ERMs in stochastic convex optimization. *arXiv preprint arXiv:2311.05398*, 2023.

[13] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021.

[14] Zixiang Chen, Junkai Zhang, Yiwen Kou, Xiangning Chen, Cho-Jui Hsieh, and Quanquan Gu. Why does sharpness-aware minimization generalize better than sgd? In *NeurIPS*, 2024.

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

[16] Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

[17] Lijun Ding, Dmitriy Drusvyatskiy, Maryam Fazel, and Zaid Harchaoui. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*, 13(2): iaae009, 2024.

[18] 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.

[19] Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent YF Tan. Efficient sharpness-aware minimization for improved training of neural networks. *arXiv preprint arXiv:2110.03141*, 2021.

[20] Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. *Advances in Neural Information Processing Systems*, 35:23439–23451, 2022.

[21] Gintare Karolina Dziugaite and Daniel Roy. Entropy-sgd optimizes the prior of a pac-bayes bound: Generalization properties of entropy-sgd and data-dependent priors. In *ICML*, 2018.

[22] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[23] Itay Evron, Ran Levinstein, Matan Schliserman, Uri Sherman, Tomer Koren, Daniel Soudry, and Nathan Srebro. Better rates for random task orderings in continual linear models. *arXiv preprint arXiv:2504.04579*, 2025.

[24] Vitaly Feldman. Generalization of ERM in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

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

[26] Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357. PMLR, 2021.

[27] Moritz Hardt, Ben Recht, and Yoram Singer. Train Faster, Generalize Better: Stability of Stochastic Gradient Descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

[28] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[29] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three Factors Influencing Minima in SGD. In *International Conference of Artificial Neural Networks (ICANN)*, 2018.

[30] Xiaowen Jiang and Sebastian U Stich. Adaptive SGD with polyak stepsize and line-search: Robust convergence and variance reduction. In *NeurIPS*, 2023.

[31] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *ICLR*, 2019.

[32] Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt J Kusner. When do flat minima optimizers work? *Advances in Neural Information Processing Systems*, 35:16577–16595, 2022.

[33] 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 *ICLR*, 2016.

[34] Hoki Kim, Jinseong Park, Yujin Choi, and Jaewook Lee. Stability analysis of sharpness-aware minimization. *arXiv preprint arXiv:2301.06308*, 2023.

[35] Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher sam: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pages 11148–11161. PMLR, 2022.

[36] Tomer Koren, Roi Livni, Yishay Mansour, and Uri Sherman. Benign underfitting of stochastic gradient descent. *Advances in Neural Information Processing Systems*, 35:19605–19617, 2022.

[37] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on machine learning*, pages 5905–5914. PMLR, 2021.

[38] Hojoon Lee, Hanseul Cho, Hyunseung Kim, Daehoon Gwak, Joonkee Kim, Jaegul Choo, Se-Young Yun, and Chulhee Yun. Plastic: Improving input and label plasticity for sample efficient reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 62270–62295, 2023.

[39] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

[40] Bingcong Li and Georgios Giannakis. Enhancing sharpness-aware optimization through variance suppression. In *NeurIPS*, 2023.

[41] Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. Friendly sharpness-aware minimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5631–5640, 2024.

[42] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021.

[43] Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12360–12370, 2022.

[44] Roi Livni. The sample complexity of gradient descent in stochastic convex optimization. *arXiv preprint arXiv:2404.04931*, 2024.

[45] Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35:34689–34708, 2022.

[46] Chao Ma and Lexing Ying. On linear stability of sgd and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34:16805–16817, 2021.

[47] Peng Mi, Li Shen, Tianhe Ren, Yiyi Zhou, Xiaoshuai Sun, Rongrong Ji, and Dacheng Tao. Make sharpness-aware minimization stronger: A sparsified perturbation approach. In *NeurIPS*, 2022.

[48] Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pages 16270–16295. PMLR, 2022.

[49] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[50] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[51] Konstantinos E Nikolakakis, Farzin Haddadpour, Amin Karbasi, and Dionysios S Kalogerias. Beyond lipschitz: Sharp generalization and excess risk bounds for full-batch gd. *arXiv preprint arXiv:2204.12446*, 2022.

[52] Matthew D Norton and Johannes O Royset. Diametrical risk minimization: Theory and computations. *Machine Learning*, 112(8):2933–2951, 2023.

[53] Sameera Ramasinghe, Lachlan Ewen MacDonald, Moshiur Farazi, Hemanth Saratchandran, and Simon Lucey. How much does initialization affect generalization? In *International Conference on Machine Learning*, pages 28637–28655. PMLR, 2023.

[54] Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3380–3394. PMLR, 02–05 Jul 2022.

[55] Matan Schliserman, Uri Sherman, and Tomer Koren. The dimension strikes back with gradients: Generalization of gradient methods in stochastic convex optimization. In *Algorithmic Learning Theory*, pages 1041–1107. PMLR, 2025.

[56] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

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

[58] Sidak Pal Singh, Hossein Mobahi, Atish Agarwala, and Yann Dauphin. Avoiding spurious sharpness minimization broadens applicability of sam. *arXiv preprint arXiv:2502.02407*, 2025.

[59] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

[60] Hao Sun, Li Shen, Qihuang Zhong, Liang Ding, Shixiang Chen, Jingwei Sun, Jing Li, Guangzhong Sun, and Dacheng Tao. Adasam: Boosting sharpness-aware minimization with adaptive learning rate and momentum for training deep neural networks. *Neural Networks*, 169:506–519, 2024. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2023.10.044.

[61] Behrooz Tahmasebi, Ashkan Soleymani, Dara Bahri, Stefanie Jegelka, and Patrick Jaillet. A universal class of sharpness-aware minimization algorithms. In *ICML*, 2024.

[62] Chengli Tan, Jiangshe Zhang, Junmin Liu, Yicheng Wang, and Yunda Hao. Stabilizing sharpness-aware minimization through a simple renormalization strategy. *Journal of Machine Learning Research*, 26(68):1–35, 2025.

[63] Shira Vansover-Hager, Tomer Koren, and Roi Livni. Rapid overfitting of multi-pass stochastic gradient descent in stochastic convex optimization. *arXiv preprint arXiv:2505.08306*, 2025.

[64] Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. *Advances in neural information processing systems*, 32, 2019.

[65] Colin Wei and Tengyu Ma. Improved sample complexities for deep networks and robust classification via an all-layer margin. *arXiv preprint arXiv:1910.04284*, 2019.

[66] Colin Wei, Yining Chen, and Tengyu Ma. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Advances in Neural Information Processing Systems*, 35:12071–12083, 2022.

[67] Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How does sharpness-aware minimization minimize sharpness? *arXiv preprint arXiv:2211.05729*, 2022.

[68] Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. *Advances in Neural Information Processing Systems*, 36:1024–1035, 2023.

[69] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *NeurIPS*, 2020.

[70] Lei Wu and Weijie J Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *International Conference on Machine Learning*, pages 37656–37684. PMLR, 2023.

[71] Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

[72] Wanyun Xie, Thomas Pethick, and Volkan Cevher. Sampa: Sharpness-aware minimization parallelized. In *NeurIPS*, 2024.

[73] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[74] Yang Zhao, Hao Zhang, and Xiuyuan Hu. Randomized sharpness-aware training for boosting computational efficiency in deep learning. *arXiv preprint arXiv:2203.09962*, 2022.

[75] Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8156–8165, 2021.

[76] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C. Dvornek, Sekhar Tatikonda, James S. Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *ICLR*, 2022.

Appendix A. Additional related work

Flat minima and generalization. The conjectured connection between flat minima and generalization dates back to Hochreiter and Schmidhuber [28]. Since then, a large body of empirical and theoretical work has suggested that flatter minima correlate with, or even guarantee, better generalization performance [10, 16, 17, 22, 25, 26, 29, 31, 33, 42, 45, 46, 48, 50, 52, 64–66, 70, 71]. However, several works caution against interpreting flatness as a universal predictor of generalization [5, 18, 53, 68]. Notably, from a theoretical perspective, Dinh et al. [18] showed that in ReLU networks sharpness can be arbitrarily altered through reparameterization without affecting the learned function or its generalization, implying that common flatness measures are not parameterization-invariant and may therefore be misleading. More recently, Wen et al. [68] examined two-layer ReLU networks defining flatness as the trace of the Hessian. Using this architecture and notion of flatness they identified scenarios where flat minima fail to generalize, while sharpness-minimization algorithms such as SAM may still succeed, although their analysis of SAM was only empirical. Our results go beyond both works: unlike [18], we give explicit constructions where flat minimizers fail while sharp minimizers generalize perfectly, directly challenging the conjecture itself, and unlike [68], we establish this phenomenon already in the fundamental convex β-smooth setting and under much stronger flatness assumptions. Furthermore we provide theoretically provable lower bounds on the generalization of SAM, offering a more rigorous understanding of its limitations.

Convergence rates of SAM. Many works on the convergence of SAM analyze a variant of SAM that does not use gradient normalization during the ascent step [1, 3, 9, 34]. This variant does not match practical implementations of SAM, where normalization is typically used [57], and more recent work showed that normalization improves SAM's performance [15]. Our work considers SAM with normalization and provides more practical bounds. Another line of research studies the implicit bias of SAM and its variants under infinitesimal step sizes [3, 67], while we focus on the practical discrete setting. In more specific cases, Bartlett et al. [8] gave convergence rates for SAM on convex quadratics, whereas our work addresses general smooth convex objectives. Recent works also consider smooth nonconvex objectives with decaying or sufficiently small r [47, 60, 76], but such assumptions differ from practice, where r might be a constant. Our bounds instead cover smooth convex functions and hold for any r, including large values. Finally, Si and Yun [57] derived convergence guarantees in deterministic and stochastic regimes, but in the smooth convex case they only proved convergence to stationary points, leaving convergence to global minima as an open problem. We close this gap by providing the first rates of convergence to global minima for SAM on general smooth convex objectives, and we are the first to incorporate the true flatness of the objective into the convergence analysis.

Generalization of SAM. Foret et al. [25], who originally introduced SAM, established PAC-Bayes bounds to explain its generalization. These bounds are dimension dependent and may be vacuous in many scenarios. More recently, Tan et al. [62] analyzed the smooth and strongly convex setting, comparing the algorithmic stability of SAM and SGD. Chen et al. [14] studied generalization from a different angle, comparing the conditions for benign overfitting under SGD and SAM in two-layer convolutional ReLU networks. In contrast to these works, we establish the first dimension-independent generalization bounds for the broad class of smooth convex (but not strongly convex) objectives, together with the first lower bounds on the generalization performance of SAM in this setting.

Generalization in SCO. Stochastic convex optimization is a fundamental theoretical framework for analyzing widely used optimization algorithms, where the loss function is assumed to be convex and Lipschitz. In this setting, prior work [12, 24, 56] have shown that, although learning in this framework is possible (e.g., via Stochastic Gradient Descent), empirical risk minimization (ERM) may fail (even under additional assumptions such as smoothness and realizability), since uniform convergence does not generally hold. In our work, we focus on flat ERMs, namely minimizers of the SAER, and demonstrate that even when the minima are flat, they may still generalize poorly. Beyond ERM, several natural algorithms such as full-batch Gradient Descent and multi-pass Stochastic Gradient Descent have also been shown to fail in this setting [2, 44, 55, 63]. All of these works focus on the non-smooth regime and establish lower bounds in that setting. In contrast, our work studies the generalization of Sharpness-Aware Minimization algorithms in smooth and realizable SCO, and we show that even under these strong assumptions, SA-GD and SAM may still generalize poorly.

Smooth SCO with low noise. The problem of smooth stochastic convex optimization with low noise as been extensively studied. [59] established that Stochastic Gradient Descent (SGD) attains a risk bound of $O\left(1/n\right)$ in this setting. This result was recently extended by [6] to the last iterate of SGD. In our work, we demonstrate that in the deterministic setting, SA-GD and SAM also attain these optimal rates when applied to smooth loss functions. In addition, for SA-GD we prove an even stronger result: under an additional flatness condition, the method achieves the same fast rates for convergence with respect to the SAER F_S^r, a function that is generally non-smooth. From a generalization perspective, recent work [6, 23, 39, 51, 54] has used stability arguments to show that gradient methods such as GD and SGD, both with and without replacement and with $T = n$, achieve an optimal risk of $O(1/n)$ in this setting. Our work shows that, in contrast to those algorithms, SA-GD and SAM may generalize poorly, even in smooth and realizable SCO.

Appendix B. Optimization Error of SA-GD and SAM

In this section we analyze the optimization performance of SA-GD and SAM. In particular, we establish bounds on both the empirical risk F_S and the sharpness-aware empirical risk F_S^r for the models output by these algorithms.

B.1. Optimization error of SA-GD

We begin with the following theorem on the optimization error of SA-GD. This result highlights the role of flatness in the convergence rate of the algorithm. When the radius of flatness ρ is small, the algorithm exhibits an additive term of $O(r^2)$ in the bound on the SAER. In contrast, when ρ is large, even with $r \approx \rho$, SA-GD minimizes the SAER and converges to a flat minimum of the empirical risk. Moreover, although SA-GD can be viewed as GD applied to a potentially non-smooth function,[4] in this latter case its convergence rate matches that of GD on smooth functions. The result is formalized in the following theorem.

Theorem 4 *For every z, assume that $f(w, z)$ is β-smooth, convex, non-negative and ρ-flat. Let $\{w_t\}_{t=1}^T$ be produced by SA-GD for T steps (Eq. (4)) with $\eta \leq 1/4\beta$ and $r > 0$. For $\widehat{w} \coloneqq \frac{1}{T}\sum_{t=1}^{T}, w_t$*

[4] For example, if $F_S = x^2$, which is β-smooth for $\beta = 2$, $F_S^r = (|x| + r)^2$, which is a non smooth function.

it holds that

$$F_S\left(\widehat{w}\right) \le F_S^r\left(\widehat{w}\right) \le \|w_1 - w^\star\|^2/\eta T + 4\beta \max\{r-\rho, 0\}^2.$$

In particular, when $\eta = \frac{1}{4\beta}$, $\|w_1 - w^\star\| = O(1)$ *and* $r - \rho = O\left(\frac{1}{\sqrt{T}}\right)$, *it holds that*

$$F_S\left(\widehat{w}\right) \le F_S^r\left(\widehat{w}\right) \le O\left(\beta/T\right).$$

For the proof, we first make use of the following key lemma, which establishes a regret bound for general algorithms whose update rule takes the form $w_{t+1} = w_t - \eta \nabla F_S(w_t + v_t)$, for $\|v_t\| \le r$.

Lemma 1 *Assume that for every* z, $f(w,z)$ *is* β*-smooth, convex, non-negative and* ρ*-flat. Let* A *be an algorithm that given a data set* S*, produces a sequence* $\{w_t\}_{t=1}^T$ *such that* $w_{t+1} = w_t - \eta \nabla F_S(w_t + v_t)$*, where* $\{v_t\}_{t=1}^T$ *are vectors such that for every* t, $\|v_t\| \le r$ *and* $\eta \le 1/4\beta$*. It holds that,*

$$\frac{1}{T}\sum_{i=1}^{T} F_S\left(w_t + v_t\right) - F_S(w^\star) \le \frac{\|w_1 - w^\star\|^2}{\eta T} + 4\beta \max\{r-\rho, 0\}^2.$$

B.2. Optimization error of SAM

For SAM, we establish the following bound on the empirical risk. Similarly to SA-GD, the flatness of the empirical risk plays a significant role in the convergence of SAM, achieving fast convergence rates the function is ρ-flat and $r \le \rho + \frac{1}{\sqrt{T}}$.

Theorem 5 *For every* z*, assume that* $f(w,z)$ *is* β*-smooth, convex, non-negative and* ρ*-flat. Let* $\{w_t\}_{t=1}^T$ *be produced by SAM for* T *steps (Eq. (4)) with* $\eta \le 1/4\beta$ *and* $r > 0$*. For* $\widehat{w} \coloneqq \frac{1}{T}\sum_{t=1}^T w_t$*, it holds that*

$$F_S\left(\widehat{w}\right) \le \frac{\|w_1 - w^\star\|^2}{\eta T} + 4\beta \max\{r-\rho, 0\}^2.$$

In particular, if $\eta = \frac{1}{4\beta}$, $\|w_1 - w^\star\| = O(1), r - \rho = O\left(\frac{1}{\sqrt{T}}\right)$, $F_S\left(\widehat{w}\right) \le O\left(\frac{\beta}{T}\right)$.

The proof is followed by Lemma 1 and appears in Appendix B.

We note that Theorem 5 establishes convergence rates with respect to the empirical risk. A natural question is whether SAM achieves the same rates, namely $O\left(\frac{\|w_1 - w^\star\|^2}{\eta T} + \beta \max\{r-\rho, 0\}^2\right)$ for the SAER. In other words, does SAM converge to a flat minimum when setting $r \le \rho$?

In the following lemma, we show that this is not the case: SAM may incur an additional term of $\Omega(r^2)$ in the convergence rate for the SAER, even for ρ-flat functions. This demonstrates that SAM can converge to a non-flat minimum, even when a ρ-flat minimum exists.

Theorem 6 *For every* $\eta > 0, n \in \mathbb{N}, r, \rho \le \frac{1}{2}$, $W = [-1, 1]$ *then there exists an instance set* $\mathfrak{Z}$*, a loss function* $F_S : W \times \mathfrak{Z} \to \mathbb{R}$ *that is non-negative, convex,* 1*-Lipschitz,* 1*-smooth and* ρ*-flat such that SAM for* T *steps holds, for any suffix average* $\widehat{w}_\tau = \frac{1}{T-\tau+1}\sum_{t=\tau}^T w_t$*, the following while applied on* f*,*

$$\forall 0 \le r \le \frac{1}{2}, \qquad F_S^r(\widehat{w}_\tau) - F_S^r(w^\star) = \Omega(r^2),$$

That is, SAM converges to a sharp minimum.

The proof appears in Appendix B.

Appendix C. Stability and Generalization of SA-GD and SAM

In this section we show upper bounds for the population loss achieved by SA-GD and SAM. For SA-GD, we prove the following theorem,

Theorem 7 *For every z, assume that $f(w,z)$ is β-smooth, convex, non-negative and ρ-flat. Let $\{w_t\}_{t=1}^T$ be produced by SA-GD for T steps (Eq. (4)) with $\eta \leq 1/4\beta$ and $r > 0$. For $\widehat{w} \coloneqq \frac{1}{T}\sum_{t=1}^T w_t$, it holds that*

$$\mathbb{E}F(\widehat{w}) \leq O\left(\frac{\|w_1 - w^\star\|^2}{\eta T} + \left(\beta + \frac{\beta^3\eta^2T^2}{n^2}\right)\max\{r-\rho,0\}^2 + \eta\beta^2r^2T + \frac{\beta^2\eta T}{n^2}\right).$$

In particular for $T = n, \eta = O(1/\beta)$, $\|w_1 - w^\star\| = O(1)$ and $r - \rho = O\left(\frac{1}{\sqrt{T}}\right)$ it holds that,

$$\mathbb{E}F(\widehat{w}) = O\left(\frac{\beta}{n} + \beta r^2 n\right).$$

For SAM we show the following result,

Theorem 8 *For every z, assume that $f(w,z)$ is β-smooth, convex, non-negative and ρ-flat. Let $\{w_t\}_{t=1}^T$ be produced by SAM for T steps (Eq. (5)) with $\eta \leq 1/4\beta$ and $r > 0$. For $\widehat{w} \coloneqq \frac{1}{T}\sum_{t=1}^T w_t$, it holds that*

$$\mathbb{E}F(\widehat{w}) \leq O\left(\frac{\|w_1 - w^\star\|^2}{\eta T} + \left(\beta + \frac{\beta^3\eta^2T^2}{n^2}\right)\max\{r-\rho,0\}^2 + \eta\beta^2r^2T + \frac{\beta^2\eta T}{n^2}\right).$$

In particular for $T = n, \eta = O(1/\beta)$, $\|w_1 - w^\star\| = O(1)$ and $r - \rho = O\left(\frac{1}{\sqrt{T}}\right)$ it holds that,

$$\mathbb{E}F(\widehat{w}) = O\left(\frac{\beta}{n} + \beta r^2 n\right).$$

We note that for $r = 0$, the bounds in Theorems 7 and 8 coincide with the risk bounds of [39, 51] for GD and SGD in convex, smooth, realizable settings.

C.1. Stability of SA-GD and ASM

The proofs of Theorems 7 and 8 are based on algorithmic stability (e.g., [11, 27]). In this section, we revisit the main arguments required for these proofs and establish an algorithmic stability upper bound for first-order methods that minimize the SAM empirical risk. In particular, the stability bounds in this section hold for any algorithm that produces a sequence $\{w_t\}_{t=1}^T$ satisfying $w_{t+1} = w_t - \eta\nabla F_S(w_t + v_t)$, where $\{v_t\}_{t=1}^T$ is a sequence of vectors such that for every t, $\|v_t\| \leq r$ and $\eta \leq 1/(2\beta)$.

The notion of stability that we consider is on-average-leave-one-out (loo) model stability (e.g., [39, 54]). For this definition, we assume without loss of generality that there exists an example $z_0 \in \mathfrak{Z}$ for which $f(w, z_0) = 0$ for all w. (Otherwise, we can artificially augment the sample space with such an instance.) Now, given an i.i.d. sample $S = (z_1, \ldots, z_n)$, with the corresponding F_S, we define the leave-one-out samples $S^{(i)} = (z_1, \dots, z_{i-1}, z_0, z_{i+1}, \dots, z_n)$ for all $i \in [n]$, with the corresponding empirical risks:

$$\forall\, i \in [n], \qquad F_{S^{(i)}} = \frac{1}{n} \sum_{z \in S_i} f(w, z) = \frac{1}{n} \sum_{j \neq i} f(w, z_j).$$

We can now define the on-average-loo model stability for learning algorithms.

Definition 2 (ℓ_2-loo-on-average model stability) *Let $A : \mathfrak{Z}^n \to \mathbb{R}^d$ be a learning algorithm. We say that A is ℓ_2-on-average model ε-stable if for any samples S, S',*

$$\frac{1}{n} \sum_{i=1}^{n} \|A(S) - A(S^{(i)})\|^2 \leq \varepsilon. \tag{6}$$

We will denote by ε_{stab} the infimum over all ε for which Eq. (6) holds.

Previous work has shown that an ε-leave-one-out stable algorithm achieves good generalization. This is formalized in the following lemma from [54].

Lemma 2 (Lemma 7 from [54]) *Let A be an ℓ_2-on-average-loo model ε-stable learning algorithm. Then, if for every z, $f(w, z)$ is convex and β-smooth with respect to w,*

$$\mathbb{E}F(A(S)) \leq 4\mathbb{E}\big[F_S(A(S))\big] + 3\beta\varepsilon.$$

We can now state the stability upper bound that we establish. It is formalized in the following lemma,

Lemma 3 *Assume that for every z, $f(w, z)$ is β-smooth, convex, non-negative and ρ-flat. Let A be an algorithm that given a data set S, produce a sequence $\{w_t\}_{t=1}^{T}$ such that*

$$w_{t+1} = w_t - \eta \nabla F_S(w_t + v_t),$$

where $\{v_t\}_{t=1}^{T}$ are vectors such that for every t, $\|v_t\| \leq r$ and $\eta \leq 1/2\beta$. Assume that A returns the averaged iterate $\widehat{w} := \frac{1}{T} \sum_{t=1}^{T} w_t$. Then, A is ℓ_2-on-average model ε-stable with

$$\varepsilon_{stab} \leq O\left(\eta\beta r^2 T + \frac{\beta\eta T}{n^2} + \frac{\beta^2\eta^2 T^2 \max(r - \rho, 0)^2}{n^2}\right)$$

The proof of Lemma 3 appears in Appendix C. The proofs of Theorems 7 and 8 follow directly from Lemmas 2 and 3 and also appear in Appendix C.

Appendix D. Proofs for Section 3

Notations. In all of the proofs in this section, we denote by $\|\cdot\|$ the ℓ_2 norm. The symbol $\odot$ represents element-wise multiplication, i.e., $(x \odot y)(i) = x(i)\, y(i)$. Finally, we write $[x]_+$ for the element-wise ReLU function, defined as $[x]_+(i) = \max\{x(i), 0\}$.

D.1. Proof of Theorem 1

Proof of Theorem 1 Let $d = 2^n + 1$, $\mathcal{Z} = \{0,1\}^{2^n}$ and let $\mathcal{D}$ be the uniform distribution over $\mathcal{Z}$. Consider the following function:

$$f(w,z) = \frac{1}{2}\max\left\{\sqrt{\sum_{i=1}^{2^n} z(i)w(i)^2 + w(d)^2} - \rho, 0\right\}^2 .$$

We show that f is convex, 1-Lipschitz, 1-smooth and ρ-flat in Lemma 4. Since the samples are uniform over $\{0,1\}^{2^n}$ we have that for a random training set $S = \{z_1,\dots,z_n\} \overset{\text{i.i.d.}}{\sim} \mathcal{D}^n$ with probability greater than $1 - e^{-1} > \frac{1}{2}$, there exists an index $I \in [2^n]$ such that for every $z \in S,\ z(I) = 0$. For any $r \geq 0$,

$$F_S^r(w) \geq F_S(w + \text{sign}(w(d)) \cdot re_d) \geq \frac{1}{2}\max\{r - \rho, 0\}^2.$$

We will now consider $w^{(1)} = e_I$. First, from the choice of I, $F_S(e_I) = 0$. For every $\|x\| \leq r$,

$$\begin{aligned} F_S(e_I + x) &= \frac{1}{2}\max\left\{\|(e_I + x)\odot z\| - \rho, 0\right\}^2 \leq \frac{1}{2}\max\left\{\|x \odot z\| - \rho, 0\right\}^2 \\ &\leq \frac{1}{2}\max\left\{\|x\| - \rho, 0\right\}^2 \leq \frac{1}{2}\max\left\{r - \rho, 0\right\}^2, \end{aligned}$$

which shows $w^{(1)} \in \arg\max_{w\in W} F_S^r$. Finally, since with probability $\frac{1}{2}$ a new sample z' will hold $z'(I) = 1$:

$$F(e_I) - F(w^\star) \geq \frac{1}{4}\cdot(1-\rho)^2 + \frac{1}{2}\cdot 0 \geq \frac{1}{16} = \Omega(1),$$

where the last inequality holds since $\rho \leq \frac{1}{2}$. This concludes the results for $w^{(1)}$. For $w^{(2)}$ consider $w^{(2)} = \rho e_d$. For every $\delta > 0$:

$$F_S^\delta(\rho e_d) \geq F_S((\rho+\delta)e_d) = \frac{1}{2}(\rho + \delta - \rho)^2 = \frac{\delta^2}{2},$$

which shows $w^{(2)}$ is a sharp minimum. And,

$$F(\rho e_I) - F(w^\star) = \frac{1}{2}\max\{\rho - \rho\}^2 - 0 = 0.$$

which concludes the proof. ■

Lemma 4 *Fix some $z \in \mathbb{R}^{d-1}\times\{1\}$, and $\rho \geq 0$. Define the following function:*

$$\phi_z(w) = [\|\,[w\odot z]_+\,\| - \rho]_+^2,$$

then ϕ_z is convex, 1-Lipschitz, $\|z\|_\infty^2$-smooth in the unit ball, and ρ-flat.

Proof of Lemma 4 We will prove each property separately.

Convexity. Notice that $K(x,z) = \|[w\odot z]_+\|$ is convex and the function $L(x) = \max\{x-\rho, 0\}^2$ is convex and non-decreasing, hence the composition $\phi = L \circ K$ is convex.

Lipschitz continuity. We will start by computing the gradient.

$$\begin{aligned}\|\nabla\phi_z(w)\| &= \left\|\frac{\max\{\|[w\odot z]_+\| - \rho, 0\}}{\|[w\odot z]_+\|}\cdot(z\odot[w\odot z]_+)\right\| \\ &\le \|z\odot[w\odot z]_+\| \le \|z\|_\infty\|[w\odot z]_+\| \le \|z\|_\infty\|w\| \le \|z\|_\infty.\end{aligned}$$

Where the last inequality comes from the choice of W as the unit ball.

Smoothness. For $x, y \in \mathbb{R}^d$,

$$\begin{aligned}\|\nabla\phi_z(x) - \nabla\phi_z(y)\| &= \left\|z\odot\left(\frac{\left[\|x\odot z\| - \rho\right]_+}{\|x\odot z\|}x\odot z - \frac{\left[\|y\odot z\| - \rho\right]_+}{\|y\odot z\|}y\odot z\right)\right\| \\ &= \|z\|_\infty\cdot\left\|\frac{\left[\|x\odot z\| - \rho\right]_+}{\|x\odot z\|}x\odot z - \frac{\left[\|y\odot z\| - \rho\right]_+}{\|y\odot z\|}y\odot z\right\|.\end{aligned}$$

Denote

$$T(u) := \frac{[\|u\| - \rho]_+}{\|u\|}u, \qquad (T(0) := 0),$$

so the last norm is $\|T(x\odot z) - T(y\odot z)\|$. Note the identity

$$T(u) = u - \Pi_{B_\rho}(u), \qquad B_\rho := \{v : \|v\| \le \rho\},$$

Hence, using that Euclidean projection is nonexpansive and $[x]_+$ is 1-Lipschitz,

$$\begin{aligned}\|T(x\odot z) - T(y\odot z)\| &= \|\, x\odot z - \Pi_{B_\rho}(x\odot z) - (y\odot z - \Pi_{B_\rho}(y\odot z))\,\| \\ &\le \|\, x\odot z - y\odot z\,\| \\ &\le \|(x-y)\odot z\| \le \|z\|_\infty\cdot\|x-y\|.\end{aligned}$$

Combining the inequalities we showed

$$\|\nabla\phi_z(x) - \nabla\phi_z(y)\| \le \|z\|_\infty^2\|x-y\|.$$

This concludes the proof for smoothness.

Flatness. For ρ flatness we can easily see that for any $\|v\| \le \rho$ the following:

$$\phi_z(0+v,\rho) = \phi_z(v,\rho) = \frac{1}{2}\left[|[v\odot z]_+\| - \rho\right]_+^2 \le \frac{1}{2}[\|v\| - \rho]_+^2 = 0$$

It is left to show that ρ is the maximum flatness. Indeed, for every $w \in \arg\min\phi_z$:

$$\phi_z(w + \mathrm{sign}(w(d))\cdot ce_d) \ge \frac{1}{2}\max\{c-\rho, 0\}^2.$$

This implies that for $c > \rho$ we will have $\phi_z^c > 0$. ■

D.2. Proof of Theorem 2

Proof of Theorem 2 Let $\mathcal{Z} = \{-1,1\}^{2^n}$ and $\mathcal{D}$ to be the uniform distribution over $\mathcal{Z}$. For $i \in [T]$ denote $w^{(i)} = w[T \cdot (i-1)+1 : T \cdot i]$. Consider the following function:

$$f(w,z) = \frac{1}{2}\max\left\{\sqrt{\sum_{i=1}^{2^n}\sum_{j=1}^{T}\max\left\{z(i)w^{(i)}(j),0\right\}^2 + w(d)^2} - \rho, 0\right\}^2 .$$

we prove that f is convex, 1-Lipschitz, 1-smooth and ρ-flat in Lemma 4. From the definition of $\mathcal{D}$, for a sample $z \sim \mathcal{D}$ the coordinates $z(i)$ are i.i.d. uniform Bernoulli. For a random training set $S = \{z_1,\ldots,z_n\} \overset{\text{i.i.d.}}{\sim} \mathcal{D}^n$, $S \subseteq \{0,1\}^{2^n}$, we have that with probability greater than $1-e^{-1} > \frac{1}{2}$, there exists a coordinate I such that all the examples in the sample are 1 on this coordinate, that is $z(I) = 1$ for all $z \in S$. Define the following SAM-gradient-oracle:

$$O_S(w) = \frac{1}{n}\sum_{i=1}^{n}\nabla f(w + e_{I_t}, z_i),$$

for $I_t = I + t - 1$. We will prove correctness by induction. For $w_1 = 0$ for every $\|v\| \le r$ the following holds:

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n} f(0+v,z_k) &= \frac{1}{2n}\sum_{k=1}^{n}\max\left\{\sqrt{\sum_{i=1}^{2^n}\sum_{j=1}^{T}\max\left\{z_k(i)(0+v^{(i)}(j)),0\right\}^2} - \rho, 0\right\}^2 \\
&\le \frac{1}{2n}\sum_{k=1}^{n}\max\left\{\sqrt{\sum_{i=1}^{2^n}\sum_{j=1}^{T}\max\left\{v^{(i)}(j),0\right\}^2} - \rho, 0\right\}^2 \\
&\le \frac{1}{2n}\sum_{k=1}^{n}\max\left\{\|v\| - \rho, 0\right\}^2 \le \frac{1}{2}(r-\rho)^2.
\end{aligned}$$

Also for I_1 chosen by the oracle:

$$\begin{aligned}
\frac{1}{2n}\sum_{k=1}^{n} f(0+e_I,z_k) &= \frac{1}{2n}\sum_{k=1}^{n}\max\left\{\sqrt{\sum_{i=1}^{2^n}\sum_{j=1}^{T}\max\left\{z_k(i)(0+v^{(i)}(j)),0\right\}^2} - \rho, 0\right\}^2 \\
&= \frac{1}{2n}\sum_{k=1}^{n}\max\left\{\sqrt{\max\left\{0+v(I),0\right\}^2} - \rho, 0\right\}^2 = \frac{1}{2}(r-\rho)^2,
\end{aligned}$$

this concludes the base case. For the induction step we can notice that in step t it holds that $w(i) \le 0$ for every i and $w(I_t) = 0$ thus the same steps as the base case complete the proof. To see no projections take place we note that by definition:

$$O_S(w_t) = \tfrac{1}{2n}\textstyle\sum_{k=1}^{n} 2\left(\sqrt{z_k(I_t)(w_t(I_t)+r)^2} - \rho\right) \cdot \frac{z\odot([w+re_{I_t}]_+)}{\sqrt{z_k(I_t)(w_t(I_t)+r)^2}} = (r-\rho)\cdot\frac{re_{I_t}}{r} = (r-\rho)e_{I_t}.$$

This implies that at time t:

$$w_t(i) = \begin{cases} -\eta(r-\rho) & i \in \{I_j\}_{j=1}^{t-1} \\ 0 & \text{o.w.} \end{cases} .$$

Since $\eta(r-\rho) \le \frac{1}{\sqrt{T}}$, we stay inside the unit ball for the entire run of the algorithm. This dynamic also imply that for every $\tau \in [T]$ suffix average $\widehat{w}_\tau = \frac{1}{T-\tau+1}\sum_{t=\tau}^{T} w_t$ and $s \le \frac{T}{2}$ the following holds:

$$
\begin{aligned}
\widehat{w}_\tau(I_s) = \frac{1}{T-\tau+1}\sum_{t=\tau}^{T} w_t(I_s) &\le \frac{1}{T-\tau+1}\sum_{t=\max\{\tau,T/2\}}^{T} w_t(I_s) \\
&\le \frac{T-\max\{\tau,T/2\}+1}{T-\tau+1}\left(-\eta(r-\rho)\right) \le -\frac{\eta(r-\rho)}{2}.
\end{aligned}
$$

With probability $\frac{1}{2}$ a new sample z' will hold $z(I)=-1$ which gives:

$$
\begin{aligned}
F(\widehat{w}_\tau) - F(0) &\ge \frac{1}{4}\max\left\{\sqrt{\sum_{t=1}^{T}\widehat{w}_\tau^{(I)}(t)^2} - \rho, 0\right\}^2 \ge \frac{1}{4}\max\left\{\frac{\eta(r-\rho)}{2}\sqrt{\frac{T}{2}} - \rho, 0\right\}^2 \\
&\ge \frac{1}{4}\max\left\{\frac{\eta(r-\rho)}{2}\sqrt{\frac{T}{2}} - \frac{\eta(r-\rho)\sqrt{T}}{3}, 0\right\}^2 && (\rho \le r - \tfrac{3r}{3+\eta\sqrt{T}}) \\
&\ge \frac{1}{4\cdot 100^2}\eta(r-\rho)\sqrt{T} = \Omega(\eta^2(r-\rho)^2 T).
\end{aligned}
$$

■

D.3. Proof of Theorem 3

Proof of Theorem 3 Let $d = T\cdot 2^n + 1$, $\mathcal{X} = \{0,1\}^{2^n}$, $\mathcal{D}$ to be the uniform distribution over $\mathcal{X}$. Denote for every $i \in [T]$; $w^{(i)} = w[T\cdot(i-1)+1 : T\cdot i]$. Consider the following function:

$$
\begin{aligned}
f(w,z) = &\frac{1}{2}\sum_{i=1}^{2^n}\sum_{j=2}^{T} z(i) w^{(i)}(j)^2 \\
&+ \frac{1}{2}\sum_{i=1}^{2^n}\sum_{j=2}^{T}\max\left\{w^{(i)}(j) - \delta_j\left(w^{(i)}(j-1) + \lambda\cdot\mathbb{1}[j=2]\right), 0\right\}^2 \\
&+ \frac{\gamma}{2}\max\{v_z^T w + \delta_1, 0\}^2,
\end{aligned}
$$

where

$$
v_z^{(i)}(j) = \begin{cases} 0 & j \ne 1 \\ -\frac{1}{2(d-1)} & i \le 2^n,\ j=1 \text{ and } z(i)=0 \\ 1 & i \le 2^n,\ j=1 \text{ and } z(i)=1 \\ 1 & i = 2^n+1 \text{ and } j=1 \end{cases},
$$

and,

$$
\delta_1 = \frac{\eta\gamma r}{2\sqrt{d} - \eta\gamma}, \qquad \lambda = \frac{r}{4d(d-1)}, \qquad \gamma = \frac{\lambda}{\max\{1,\eta\}(r+\delta_1)}.
$$

The positive parameters $\{0 < \delta_j \le 1\}_{j=2}^{T}$ will be chosen later. We will prove f has the desired properties in the following lemma whose proof is deferred to Appendix D.3.1.

Lemma 5 *f defined as defined above is convex, 6-smooth, 7-Lipschitz and realizable, meaning ρ-flat with $\rho = 0$.*

Since the distribution $\mathcal{D}$ is uniform over $\{0,1\}^{2^n}$, for a random training set $S = \{z_1, \ldots, z_n\}$ with probability at least $\frac{1}{e} > \frac{1}{3}$, there exists *exactly one* index I such that for every $z \in S$, $z(I) = 0$. For the rest of the proof, assume this event holds. We will show the dynamics of the algorithm under this assumption in the following lemma whose proofs are deferred to Appendix D.3.1:

Lemma 6 *Assuming there exists a coordinate I such that $\forall z \in S;\ z(I) = 0$, and $\eta r \le \frac{1}{\sqrt{T}}$, $w_1 = 0$, then there exists $\delta_2 > 0$ such that after running one SAM update on F_S,*

1. *$\forall z \in S;\ v_z^T w_2 + \delta_1 \le 0$*
2. *$\forall i \ne I;\ -\lambda < w_2^{(i)}(1) < 0$*
3. *$\forall i \ne I,\ j \ge 2;\ w_2^{(i)}(j) = 0$*
4. *$\forall j \ge 3;\ w_2^{(I)}(j) = 0$*
5. *$0 \le w_2^{(I)}(1) \le \frac{1}{d}$*
6. *$-\frac{1}{d} \le w_2^{(I)}(2) < 0$.*

From this lemma we can conclude that if $w_t^{(I)}(2)$ remains negative throughout the remaining run of the algorithm, none of the coordinates in $w^{(i)}$ where $i \ne I$ will change, and neither will $w^{(i)}(1)$ for every i. This means that while $w_t^{(I)}(2)$ remains negative it suffices to prove the dynamics for the following function:

$$g(u) = \frac{1}{2}\sum_{j=3}^{T} \max\left\{u(j) - \delta_j u(j-1), 0\right\}^2 + \max\{u(2), 0\}^2,$$

when we start from $u_2 = -\sigma e_2$ for $\sigma = |w_2^{(I)}(2)| > 0$. The dynamics we will prove for $u[2:T]$ will hold for $w^{(I)}[2:T]$ while the rest of w stays the same as in w_2. We will now continue to look at the dynamics of $\{u_t\}_{t=2}^T$. We will have the following lemma whose proof is deferred to Appendix D.3.1:

Lemma 7 *There exists a set of positive parameters $\{0 < \delta_t \le 1\}_{t=3}^T$ such that starting from $u_2 = -\sigma e_2$ will give us the following for $t \ge 4$:*

1. *$-\sigma \le u_t(2) \le 0$*
2. *$u_t(i+1) - \delta_i u_t(i) \begin{cases} \le 0 & 2 \le i < t \\ > 0 & i = t \\ = 0 & t < i \le T-1 \end{cases}$*
3. *$-2\eta r \le u_t(t) \le -\eta r$*
4. *$-2\eta r \le u_t(t-1) \le -\frac{1}{2}\eta r$.*

In the proof of the dynamic of u we did not consider projections, that is because with this dynamic and the assumption that $\eta r \le \frac{1}{2\sqrt{T}}$ means we stay inside the unit ball for the entire algorithm and no projections take place. To see this notice using Lemmas 6 and 7 that for every $t \in [T]$:

$$\|w_t\|^2 \le \|w_T\|^2 \le 2(T-1) \cdot 4\eta^2 r^2 + d \cdot \frac{1}{d^2} \le 4\frac{(T-1)}{4T} + \frac{1}{T \cdot 2^n} \le 1.$$

Concluding we know that for $t = 3, \ldots, T$:

$$\forall j \in \{3, \ldots, t\};\ w_t^{(I)}(j) \leq -\frac{1}{2}\eta r.$$

This implies that for a suffix average $\tau \in [T];\ \widehat{w}_\tau = \frac{1}{T-\tau+1}\sum_{t=\tau}^{T} w_t$ we have that for $s \geq \frac{T}{2}$:

$$\begin{aligned}\widehat{w}_\tau^{(I)}(s) &= \frac{1}{T-\tau+1}\sum_{t=\tau}^{T} w_t^{(I)}(s) \leq \frac{1}{T-\tau+1}\sum_{t=\max\{\tau, T/2\}}^{T} w_t^{(I)}(s)\\ &\leq \frac{T-\max\{\tau, T/2\}+1}{T-\tau+1}\left(-\frac{1}{2}\eta r\right) \leq -\frac{\eta r}{4}.\end{aligned}$$

With probability $\frac{1}{2}$ a new sample z' will have $z'(I) = 1$. This means that for every $\tau \in [T]$:

$$F(\widehat{w}_\tau) - F(w^\star) \geq \|\widehat{w}_\tau\|^2 \geq \frac{1}{4}\eta^2 r^2 \cdot \frac{T}{2} = \Omega(\eta^2 r^2 T).$$

Where we use the fact that F is realizable. This concludes the proof. ■

D.3.1. OMITTED PROOFS

Proof of Lemma 5 We will use the following notation:

$$\begin{aligned}f(w,z) = &\underbrace{\frac{1}{2}\sum_{i=1}^{2^n}\sum_{j=2}^{T} z(i)\, w^{(i)}(j)^2}_{=:f_1(w)}\\ &+\underbrace{\frac{1}{2}\sum_{i=1}^{2^n}\sum_{j=2}^{T}\big[\, w^{(i)}(j) - \delta_j\big(w^{(i)}(j-1) + \lambda\, \mathbb{1}[j=2]\big)\,\big]_+^2}_{=:f_2(w)}\\ &+\underbrace{\frac{\gamma}{2}\,[\, v_z^\top w + \delta_1\,]_+^2,}_{=:f_3(w)}\end{aligned}$$

Convexity. Each component is convex:

- f_1: a nonnegative sum of convex quadratics.
- f_2: each term is $\frac{1}{2}(\text{affine}(w))_+^2$, convex because $x \mapsto \frac{1}{2}(x_+)^2$ is convex and nondecreasing.
- f_3: same reasoning as f_2.

Therefore f is convex.

Lipschitz continuity. We will bound the norm of the gradients inside the unite ball.

- f_1: $\nabla f_1(w) = z(i)\, w^{(i)}(j)$ on each (i,j) with $j \geq 2$, hence $\|\nabla f_1(w)\| \leq \|w\| \leq 1$.
- f_2: define $r_{i,j}(w) = \big[\, w^{(i)}(j) - \delta_j(w^{(i)}(j-1) + \lambda \mathbb{1}[j=2])\,\big]$. Each term $\frac{1}{2}r_{i,j}(w)^2$ contributes gradient supported on $w^{(i)}(j), w^{(i)}(j-1)$ with squared norm $(1+\delta_j^2)r_{i,j}(w)^2 \leq 2r_{i,j}(w)^2$. Summing and bounding as in $(a-b)^2 \leq 2a^2+2b^2$, $\delta_j \leq 1$, we obtain $\|\nabla f_2(w)\| \leq 4 + 2|\lambda|\delta_2\sqrt{2^n} \leq 4 + 2\cdot\frac{r}{2d(d-1)}\sqrt{2^n} \leq 5$.

- f_3: $\nabla f_3(w) = \gamma\,(v_z^\top w + \delta_1)_+\, v_z$, hence $\|\nabla f_3(w)\|_2 \le \gamma(\|v_z\| + |\delta_1|)\|v_z\| \le \frac{1}{4d(d-1)} \cdot (\frac{d}{T} + 1) \cdot \frac{d}{T} \le 1$.

Adding the three bounds gives f is 7-Lipschitz.

Smoothness.

- f_1: its Hessian is diagonal with entries $z(i)$ on coordinates (i, j) with $j \ge 2$, hence $\|\nabla f_1(x) - \nabla f_1(y)\|_2 \le \|x - y\|_2$. So f_1 is 1-smooth.
- f_2: For each i, stack the variables as $w^{(i)} \in \mathbb{R}^T$ and define the linear map

  $$(Bw^{(i)})_{j-1} \;=\; w^{(i)}(j) - \delta_j\, w^{(i)}(j-1), \qquad j = 2, \dots, T,$$

  so $B \in \mathbb{R}^{(T-1)\times T}$ has 1 on the superdiagonal and $-\delta_j$ on the subdiagonal positions that touch it. Let $b \in \mathbb{R}^{T-1}$ encode the constant shift $b_1 = -\delta_2\,\lambda$ and $b_k = 0$ for $k \ge 2$. Writing x for the full vector that stacks all $w^{(i)}$, we can express

  $$f_2(x) = \frac{1}{2}\sum_{i=1}^{2^n} \left\| (Bw^{(i)} + b)_+ \right\|_2^2 = \frac{1}{2}\left\| (Ax + c)_+ \right\|_2^2,$$

  where A is block-diagonal with 2^n copies of B and c stacks the copies of b. Define $\phi(z) = \frac{1}{2}\|z_+\|_2^2 = \sum_k \frac{1}{2}[z_k]_+^2$. Then $\nabla\phi(z) = z_+$ and $\nabla\phi$ is 1-Lipschitz since $\|z_+ - y_+\|_2 \le \|z - y\|_2$. By the chain rule,

  $$\nabla f_2(x) = A^\top\,(Ax + c)_+ = A^\top \nabla\phi(Ax + c).$$

  Hence, for any x, y,

  $$\begin{aligned}\|\nabla f_2(x) - \nabla f_2(y)\| &= \left\|A^\top\big(\nabla\phi(Ax + c) - \nabla\phi(Ay + c)\big)\right\| \\ &\le \|A\|\,\|\nabla\phi(Ax + c) - \nabla\phi(Ay + c)\| \\ &\le \|A\|\,\|A(x - y)\| \le \|A\|^2\,\|x - y\|.\end{aligned}$$

  Therefore f_2 is $\|A\|^2 = \|B\|^2$-smooth. Using $\delta_j \le 1$ and $(a - b)^2 \le 2a^2 + 2b^2$,

  $$\|Bx\|_2^2 = \sum_{j=2}^{T}\big(x_j - \delta_j x_{j-1}\big)^2 \;\le\; 2\sum_{j=2}^{T} x_j^2 + 2\sum_{j=2}^{T}\delta_j^2 x_{j-1}^2 \;\le\; 4\sum_{j=1}^{T} x_j^2,$$

  so $\|B\| \le 2$ and consequently f_2 is 4-smooth.
- f_3: $\nabla f_3(w) = \gamma\,[\,v_z^\top w + \delta_1\,]_+\, v_z$. For any x, y,

  $$\|\nabla f_3(x) - \nabla f_3(y)\| = \gamma\left|[\,v_z^\top x + \delta_1\,]_+ - [\,v_z^\top y + \delta_1\,]_+\right|\,\|v_z\| \le \gamma\,|v_z^\top(x - y)|\,\|v_z\| \le \gamma\|v_z\|^2\|x - y\|,$$

  so f_3 is $\gamma\|v_z\|^2$-smooth.

Summing gives f is $\beta \le 5 + \gamma\|v_z\|_2^2 \le 5 + d \cdot \frac{1}{4d(d-1)} \le 6$ smooth.

Realizability. We can see that for

$$w^\star(i) = \begin{cases} 0 & i < d \\ -\frac{\lambda}{2} & i = d \end{cases}$$

$f(w^\star, z) = 0$ for every $z \in \{0,1\}^d$. ■

Proof of Lemma 6 Denote $v_S = \frac{1}{n}\sum_{k=1}^n v_{z_k}$. We will compute the gradient steps explicitly,

$$\nabla F(w_1) = \frac{1}{n}\sum_{k=1}^n \delta_1 \cdot \gamma v_{z_k} = \delta_1 \gamma v_S.$$

Hence,

$$w_{1+1/2} = 0 + \frac{r\delta_1\gamma v_S}{\delta_1\gamma\|v_S\|} = \frac{r v_S}{\|v_S\|}.$$

Since $n \geq 2$ it holds that $\frac{1}{2(d-1)} \leq \frac{1}{2n}$. This implies $v_z(i) = 1 \implies w_{1+1/2}(i) > 0$. Hence, for every $z \in S$:

$$v_z^T w_{1+1/2} \geq w_1^{(2^n+1)}(1) - \frac{r}{2(d-1)} \cdot \frac{d-1}{\|v_S\|} = \frac{r}{2\|v_S\|} > 0.$$

We can calculate the first SAM update explicitly,

$$\begin{aligned}
\nabla F_S(w_{1+1/2}) &= \frac{1}{n}\sum_{k=1}^n \gamma\left(v_{z_k} \odot \frac{r v_S}{\|v_S\|} + \delta_1\right) \odot v_{z_k} + \left[-\delta_2\left(\frac{r v_S(I)}{\|v_S\|} + \lambda\right)\right]_+ (e_2^{(I)} - \delta_2 e_1^{(I)}) \\
&= \gamma\left(\frac{r v_S}{\|v_S\|}\left[\frac{1}{n}\sum_{k=1}^n v_{z_k} \odot v_{z_k}\right] + \delta_1 v_S\right) - \delta_2\left(\frac{r v_S(I)}{\|v_S\|} + \lambda\right)(e_2^{(I)} - \delta_2 e_1^{(I)}),
\end{aligned}$$

where the last step is from the fact that:

$$\frac{v_S(I)}{\|v_S\|} + \lambda = -\frac{r}{2(d-1)\|v_S\|} + \lambda \leq -\frac{1}{2(d-1)d} + \frac{r}{4d(d-1)} = -\frac{r}{4d(d-1)} < 0.$$

Notice that similarly to before, this gradient step guarantees $v_z(i) = 1 \implies w_2(i) < 0$. Since $v_S(T \cdot 2^n + 1) = 1$, for every $z \in S$:

$$\begin{aligned}
v_z^T w_2 &\leq w_2^{(2^n+1)}(1)\left(1 - \frac{1}{2(d-1)}(d-1)\right) = \frac{1}{2} w_2^{(2^n+1)}(1) = -\frac{\eta\gamma}{2}\left(\frac{r+\delta_1}{\|v_S\|}\right) \\
&\leq -\frac{\eta\gamma(r+\delta_1)}{2\sqrt{d}} < 0 && (\|v_S\| \leq \sqrt{d})
\end{aligned}$$

This implies that for every $z \in S$:

$$v_z^T w_2 + \delta_1 \leq -\frac{\eta\gamma(r+\delta_1)}{2\sqrt{d}} + \delta_1 = \frac{-\eta\gamma r + \delta_1(2\sqrt{d} - \eta\gamma)}{2\sqrt{d}} = 0.$$

Where the last step is due to the choice of δ_1 and concludes Item 1. Furthermore, for every $i \neq I$ we have that:

$$w_2^{(i)}(1) + \lambda \geq -\frac{\gamma(r+\delta_1)}{\|v_S\|} + \lambda \geq -\gamma(r+\delta_1) + \lambda = 0,$$

where the last step is from the choice of γ concluding Item 2. Finally,

$$\begin{aligned}
w_2^{(I)}(1) &= -\eta\delta_2^2\left(\frac{r v_S(I)}{\|v_S\|}\right) + \eta\gamma\left(\frac{r}{8(d-1)^3\|v_S\|} + \frac{\delta_1}{2(d-1)}\right) \\
&\leq -\eta\delta_2^2\left(\frac{r v_S(I)}{\|v_S\|}\right) + \frac{1}{4d(d-1)}\left(\frac{1}{8(d-1)^3} + \frac{1}{2(d-1)}\right),
\end{aligned}$$

where the last inequality is again from the choice of γ. This implies that there exists $\tau_1 > 0$ such that for every $\delta_2 \leq \tau_1$ it holds that $0 \leq w_2^{(I)}(1) \leq \frac{1}{\sqrt{d}}$. Similarly,

$$w_2^{(I)}(2) = \eta\delta_2 \left(\frac{r v_S(I)}{\|v_S\|} + \lambda \right).$$

Since this goes to 0 when δ_2 goes to zero, there exists τ_2 such that for every $0 < \delta_2 \leq \tau_2$; $-\frac{1}{\sqrt{d}} \leq w_2^{(I)}(2) < 0$. Choosing $0 < \delta_2 = \min\{\tau_1, \tau_2, 1\}$ concludes Items 5 and 6. Items 3 and 4 hold since these coordinates weren't changed by the update and thus stayed 0. ■

Proof of Lemma 7 We will show the claim by induction on t.

Base case. We will start by computing u_4. Using all we've proved we get:

$$\nabla g(u_2) = -\delta_3\sigma(e_3 - \delta_3 e_2),$$

which gives:

$$\begin{aligned} u_{2+1/2} &= u_2 + \frac{r\nabla g(u_2)}{\|\nabla g(u_2)\|} = u_2 + \frac{r\delta_3}{\delta_3\sqrt{1+\delta_3^2}} e_3 - \frac{r\delta_3^2}{\delta_3\sqrt{1+\delta_3^2}} e_2 \\ &= u_2 + \frac{r}{\sqrt{1+\delta_3^2}} e_3 - \frac{r\delta_3}{\sqrt{1+\delta_3^2}} e_2. \end{aligned}$$

Thus,

$$\begin{aligned} &\nabla g(u_{2+1/2}) = (u_{2+1/2}(3) - \delta_3 u_{2+1/2}(2))e_3 - \delta_3(u_{2+1/2}(3) - \delta_3 u_{2+1/2}(2))e_2 \\ &= \left(\frac{r}{\sqrt{1+\delta_3^2}} + \frac{r\delta_3^2}{\sqrt{1+\delta_3^2}} - \delta_3 u_2(2) \right) e_3 - \delta_3 \left(\frac{r}{\sqrt{1+\delta_3^2}} + \frac{r\delta_3^2}{\sqrt{1+\delta_3^2}} - \delta_3 u_2(2) \right) e_2 \\ &= \left(r\sqrt{1+\delta_3^2} - \delta_3 u_2(2) \right) e_3 - \delta_3 \left(r\sqrt{1+\delta_3^2} - \delta_3 u_2(2) \right) e_2. \end{aligned}$$

Finally,

$$u_3 = u_2 - \eta \left(r\sqrt{1+\delta_3^2} - \delta_3 u_2(2) \right) e_3 + \eta\delta_3 \left(r\sqrt{1+\delta_3^2} - \delta_3 u_2(2) \right) e_2.$$

This gives:

$$-\sigma \leq u_3(2) = -\sigma + \eta\delta_3 \left(r\sqrt{1+\delta_3^2} - \delta_3 u_2(2) \right).$$

Importantly σ does not depend on δ_3 so this term goes to $-\sigma < 0$ as δ_3 goes to 0. This means that there exists τ_1 such that for every $\delta_3 \leq \tau_1$ we have that $u_3(2) < 0$. Furthermore,

$$u_3(3) = -\eta \left(r\sqrt{1+\delta_3^2} - \delta_3 u_2(2) \right) \leq -\eta r + \eta\delta_3 u_2(2) \leq -\eta r,$$

where the last inequality is from the fact that $u_2(2) \le 0$. Also since $u_3(3)$ goes to $-\eta r$ when δ_3 goes to 0, there exists τ_2 such that for $\delta_3 \le \tau_2$:

$$u_3(3) = -\eta\left(r\sqrt{1+\delta_3^2} - \delta_3 u_2(2)\right) \ge -2\eta r.$$

Further,

$$u_3(3) - \delta_3 u_3(2) = -\eta\left(r\sqrt{1+\delta_3^2} - \delta_3 u_2(2)\right) - \delta_3\left(-\sigma + \eta\delta_3\left(r\sqrt{1+\delta_3^2} - \delta_3 u_2(2)\right)\right).$$

Again, this term goes to something strictly negative as δ_3 goes to 0. This means that there exists τ_3 such that for every $\delta_3 \le \tau_3$ it holds that $u_3(3) - \delta_3 u_3(2) < 0$. Choosing $\delta_3 = \min\{\tau_1, \tau_2, \tau_3, 1\}$ concludes u_3. We will now calculate u_4. From what we have shown:

$$\nabla g(u_3) = -\delta_4 u_3(3)(e_4 - \delta_4 e_3),$$

which gives:

$$\begin{aligned} u_{3+1/2} &= u_3 + \frac{r\nabla g(u_3)}{\|\nabla g(u_3)\|} = u_3 + \frac{r\delta_4}{\delta_4\sqrt{1+\delta_4^2}}e_4 - \frac{r\delta_4^2}{\delta_4\sqrt{1+\delta_4^2}}e_3 \\ &= u_3 + \frac{r}{\sqrt{1+\delta_4^2}}e_4 - \frac{r\delta_4}{\sqrt{1+\delta_4^2}}e_3. \end{aligned}$$

Thus,

$$\begin{aligned} &\nabla g(u_{3+1/2}) = (u_{3+1/2}(4) - \delta_4 u_{3+1/2}(3))e_4 - \delta_4(u_{3+1/2}(4) - \delta_4 u_{3+1/2}(3))e_3 \\ &= \left(\frac{r}{\sqrt{1+\delta_4^2}} + \frac{r\delta_4^2}{\sqrt{1+\delta_4^2}} - \delta_4 u_3(3)\right)e_4 - \delta_4\left(\frac{r}{\sqrt{1+\delta_4^2}} + \frac{r\delta_4^2}{\sqrt{1+\delta_4^2}} - \delta_4 u_3(3)\right)e_3 \\ &= \left(r\sqrt{1+\delta_4^2} - \delta_4 u_3(3)\right)e_4 - \delta_4\left(r\sqrt{1+\delta_4^2} - \delta_4 u_3(3)\right)e_3. \end{aligned}$$

Finally,

$$u_4 = u_3 - \eta\left(r\sqrt{1+\delta_4^2} - \delta_4 u_3(3)\right)e_4 + \eta\delta_4\left(r\sqrt{1+\delta_4^2} - \delta_4 u_3(3)\right)e_3.$$

This gives:

$$-2\eta r \le u_4(3) = u_3(3) + \eta\delta_4\left(r\sqrt{1+\delta_4^2} - \delta_4 u_3(3)\right).$$

Importantly $u_3(3)$ does not depend on δ_4 so this term goes to $u_3(3) < -\eta r$ as δ_4 goes to 0. This means that there exists θ_1 such that for every $\delta_4 \le \theta_1$ we have that $u_4(3) < -\frac{1}{2}\eta r$. Furthermore,

$$u_4(4) = -\eta\left(r\sqrt{1+\delta_4^2} - \delta_4 u_3(3)\right) \le -\eta r + \eta\delta_4 u_3(3) \le -\eta r,$$

where the last inequality is from the fact that $u_3(3) \leq 0$. Also since $u_4(4)$ goes to $-\eta r$ when δ_4 goes to 0, there exists θ_2 such that for $\delta_4 \leq \theta_2$:

$$u_4(4) = -\eta \left(r\sqrt{1+\delta_4^2} - \delta_4 u_3(3) \right) \geq -2\eta r.$$

Further,

$$u_4(4) - \delta_4 u_4(3) = -\eta \left(r\sqrt{1+\delta_4^2} - \delta_4 u_3(3) \right) - \delta_4 \left(-u_3(3) + \eta\delta_4 \left(r\sqrt{1+\delta_4^2} - \delta_4 u_3(3) \right) \right).$$

Again, this term goes to something strictly negative as δ_4 goes to 0. This means that there exists θ_3 such that for every $\delta_4 \leq \theta_3$ it holds that $u_4(4) - \delta_4 u_4(3) < 0$. Choosing $\delta_4 = \min\{\theta_1, \theta_2, \theta_3, 1\}$ concludes u_4 and the base case.

Inductive step. Assume this holds for $t' \leq t$. Notice that from the claim it holds that $u_{t'}$ does not depend on δ_t for $t' \leq t$. So we can choose δ_t now using $\{u_{t'}\}_{t' \leq t}$. We will calculate the SAM update for from u_{t-1} to u_t using the inductive assumption:

$$\nabla g(u_{t-1}) = -\delta_t u_t(t)(e_t - \delta_t e_{t-1})$$

which gives:

$$\begin{aligned} u_{t-1+1/2} = u_{t-1} + \frac{r\nabla g(u_{t-1})}{\|\nabla g(u_{t-1})\|} &= u_{t-1} + \frac{r\delta_t}{\delta_t\sqrt{1+\delta_t^2}} e_t - \frac{r\delta_t^2}{\delta_t\sqrt{1+\delta_t^2}} e_{t-1} \\ &= u_{t-1} + \frac{r}{\sqrt{1+\delta_t^2}} e_t - \frac{r\delta_t}{\sqrt{1+\delta_t^2}} e_{t-1}. \end{aligned}$$

Thus,

$$\begin{aligned} \nabla g(u_{t-1+1/2}) &= (u_{t-1+1/2}(t) - \delta_t u_{t-1+1/2}(t-1))(e_t - \delta_t e_{t-1}) \\ &= \left(\frac{r}{\sqrt{1+\delta_t^2}} + \frac{r\delta_t^2}{\sqrt{1+\delta_t^2}} - \delta_t u_{t-1}(t-1) \right)(e_t - \delta_t e_{t-1}) \\ &= \left(r\sqrt{1+\delta_t^2} - \delta_t u_{t-1}(t-1) \right) e_t - \delta_t \left(r\sqrt{1+\delta_t^2} - \delta_t u_{t-1}(t-1) \right) e_{t-1}. \end{aligned}$$

Finally,

$$u_t = u_{t-1} - \eta \left(r\sqrt{1+\delta_t^2} - \delta_t u_{t-1}(t-1) \right) e_t + \eta\delta_t \left(r\sqrt{1+\delta_t^2} - \delta_t u_3 t - 1(t-1) \right) e_{t-1}.$$

This gives:

$$-2\eta r \leq u_t(t-1) = u_{t-1}(t-1) + \eta\delta_t \left(r\sqrt{1+\delta_t^2} - \delta_t u_{t-1}(t-1) \right).$$

Importantly $u_{t-1}(t-1)$ does not depend on δ_t so this term goes to $u_{t-1}(t-1) < -\eta r$ as δ_t goes to 0. This means that there exists θ_1 such that for every $\delta_t \leq \theta_1$ we have that $u_t(t-1) < -\frac{1}{2}\eta r$. Furthermore,

$$u_t(t) = -\eta\left(r\sqrt{1+\delta_t^2} - \delta_t u_{t-1}(t-1)\right) \leq -\eta r + \eta\delta_t u_{t-1}(t-1) \leq -\eta r,$$

where the last inequality is from the fact that $u_{t-1}(t-1) \leq 0$. Also since $u_t(t)$ goes to $-\eta r$ when δ_4 goes to 0, there exists θ_2 such that for $\delta_t \leq \theta_2$:

$$u_t(t) = -\eta\left(r\sqrt{1+\delta_t^2} - \delta_t u_{t-1}(t-1)\right) \geq -2\eta r.$$

Further,

$$\begin{aligned}&u_t(t) - \delta_t u_t(t-1) = \\ &-\eta\left(r\sqrt{1+\delta_t^2} - \delta_t u_{t-1}(t-1)\right) - \delta_t\left(-u_{t-1}(t-1) + \eta\delta_t\left(r\sqrt{1+\delta_t^2} - \delta_t u_{t-1}(t-1)\right)\right).\end{aligned}$$

Again, this term goes to something strictly negative as δ_t goes to 0. This means that there exists θ_3 such that for every $\delta_t \leq \theta_3$ it holds that $u_t(t) - \delta_t u_t(t-1) < 0$. Choosing $\delta_t = \min\{\theta_1, \theta_2, \theta_3\}$ concludes u_t and the proof. ■

Appendix E. Proofs for Appendix B

In the proofs, we use the following standard lemma (e.g., [59]).

Lemma 8 *For a non-negative and β-smooth $f : \mathbb{R}^d \to \mathbb{R}$, it holds that $\|\nabla f(w)\|^2 \leq 2\beta f(w)$ for all $w \in \mathbb{R}^d$.*

Proof of Lemma 1 By Definition 1 we know that there exists a model $w^\star$ such that for every $\|v\| \leq \rho$, it holds that $F_S(w^\star) = F_S(w^\star + v) = 0$. By Lemma 8 and Young's inequality, since for every t, we know that $w_{t+1} = w_t - \eta\nabla F_S(w_t + v_t)$, it holds for every $\gamma > 0$ that,

$$\begin{aligned}\|w_{t+1} - w^\star\|^2 &\leq \|w_t - w^\star\|^2 - 2\langle\eta\nabla F_S(w_t+v_t), w_t - w^\star\rangle + \eta^2\|\nabla F_S(w_t+v_t)\|^2 \\ &\leq \|w_t - w^\star\|^2 - 2\left\langle\eta\nabla F_S(w_t+v_t), w_t + v_t - w^\star - \min\{\rho, \|v_t\|\}\frac{v_t}{\|v_t\|}\right\rangle \\ &\quad + 2\left\langle\eta\nabla F_S(w_t+v_t), v_t - \min\{\rho, \|v_t\|\}\frac{v_t}{\|v_t\|}\right\rangle + 2\eta^2\beta F_S(w_t+v_t) - 2\eta^2\beta F_S(w^\star) \\ &\leq \|w_t - w^\star\|^2 - 2\eta F_S(w_t+v_t) + 2\eta F_S(w^\star) + \frac{1}{\gamma}\eta^2\|\nabla F_S(w_t+v_t)\|^2 + \gamma\max\{r-\rho, 0\}^2 \\ &\quad + 2\eta^2\beta F_S(w_t+v_t) - 2\eta^2\beta F_S(w^\star).\end{aligned}$$

For $\gamma = 4\eta\beta$ and $\eta \leq \frac{1}{4\beta}$, we get that,

$$\begin{aligned}
\|w_{t+1} - w^\star\|^2 \leq{}& \|w_t - w^\star\|^2 - 2\eta F_S(w_t + v_t) + 2\eta F_S(w^\star) + \frac{\eta}{4\beta}\|\nabla F_S(w_t + v_t)\|^2 \\
&+ 4\eta\beta \max\{r - \rho, 0\}^2 + 2\eta^2\beta F_S(w_t + v_t) - 2\eta^2\beta F_S(w^\star) \\
\leq{}& \|w_t - w^\star\|^2 - 2\eta F_S(w_t + v_t) + 2\eta F_S(w^\star) + \frac{\eta}{2}F_S(w_t + v_t) - \frac{\eta}{2}F_S(w^\star) + \\
&4\eta\beta \max\{r - \rho, 0\}^2 + 2\eta^2\beta F_S(w_t + v_t) - 2\eta^2\beta F_S(w^\star) \qquad \text{(Lemma 8)} \\
\leq{}& \|w_t - w^\star\|^2 + 4\eta\beta \max\{r - \rho, 0\}^2 - \eta F_S(w_t + v_t) + \eta F_S(w^\star)
\end{aligned}$$

Averaging from 1 to T and rearragining, we get the lemma. ■

Proof of Theorem 4 Let $\bar{v} = \arg\max_{\|v\| \leq r} F_S(v + \frac{1}{T}\sum_{t=1}^T w_t)$, thus, by Lemma 1, using Jensen inequality, we get

$$\begin{aligned}
F_S^r\left(\frac{1}{T}\sum_{t=1}^T w_t\right) &= F_S^r\left(\frac{1}{T}\sum_{t=1}^T w_t\right) - F_S(w^\star) \\
&= F_S\left(\frac{1}{T}\sum_{t=1}^T w_t + \bar{v}\right) - F_S(w^\star) \\
&\leq \frac{1}{T}\sum_{i=1}^T F_S\left(w_t + \bar{v}\right) - F_S(w^\star) \\
&\leq \frac{1}{T}\sum_{i=1}^T F_S\left(w_t + v_t\right) - F_S(w^\star) \\
&\leq \frac{\|w_1 - w^\star\|^2}{\eta T} + 4\beta \max\{r - \rho, 0\}^2.
\end{aligned}$$

■

Proof of Theorem 5 By the convexity of F_S we know that, for every t,

$$\begin{aligned}
F_S(w_t + v_t) &\geq F_S(w_t) + \langle \nabla F_S(w_t), v_t \rangle \\
&= F_S(w_t) + \langle \nabla F_S(w_t), r\frac{\nabla F_S(w_t)}{\|\nabla F_S(w_t)\|} \rangle \\
&= F_S(w_t) + r\|\nabla F_S(w_t)\| \\
&\geq F_S(w_t).
\end{aligned}$$

Then, by Lemma 1, using Jensen inequality, we get,

$$\begin{aligned}
F_S\left(\frac{1}{T}\sum_{t=1}^T w_t\right) &\leq \frac{1}{T}\sum_{i=1}^T F_S\left(w_t\right) - F_S(w^\star) \\
&\leq \frac{1}{T}\sum_{i=1}^T F_S\left(w_t + v_t\right) - F_S(w^\star) \\
&\leq \frac{\|w_1 - w^\star\|^2}{\eta T} + 4\beta \max\{r - \rho, 0\}^2.
\end{aligned}$$

■

Proof of Theorem 6 Let $f(w) = \frac{1}{2}\max(0,x)^2$. Its (one-dimensional) derivatives are, for $w \neq 0$,

$$f'(w) = w, \qquad f''(w) = 1,$$

and for $w < 0$,

$$f'(w) = f''(w) = 0,$$

f is a non-negative function. The convexity is implied by the positivity of f''. The Lipschitzness is implied by the fact that $|f'(w)| \leq 1$ for every $w \in W$. The smoothness is followed by the fact that $g(w) = \max(0,w)$ is a Lipschitz function as a max function over two Lipschitz functions. In addition, f is ρ-flat since $w^\star = -\frac{1}{2}$, holds $f(w^\star + v) = 0$ for every $\|v\| \leq \frac{1}{2}$. Now, let $w_1 = 0$. Since $f'(0) = 0$, $w_2 = w_1 = 0$ and by induction it follows that SAM satisfies $w_t = 0$ for every t. As a result, for any τ, $\widehat{w}_\tau = 0$, and, for every $0 \leq r \leq \frac{1}{2}$, it holds that,

$$F_S^r(\widehat{w}_\tau) - F_S^r(w^\star) = \max_{v \leq r} \frac{1}{2}\max(0,v)^2 - 0 = \frac{1}{2}r^2.$$

■

Appendix F. Proofs for Appendix C

Proof of Lemma 3 Denote by $\{w_t^{(i)}\}_{t\in[T]}$ the iterates of $S^{(i)}$ and by $\{v_t^{(i)}\}_{t\in[T]}$ the corresponding sequence of perturbations vectors. It holds that,

$$\begin{aligned}\|w_{t+1} - w_{t+1}^{(i)}\|^2 &= \|w_t - w_t^{(i)} - \eta\big(\nabla F_S(w_t + v_t) - \nabla F_{S^{(i)}}(w_t^{(i)} + v_t^{(i)})\big)\|^2 \\ &\leq \delta_t^2 + \underbrace{\eta^2\|\nabla F_S(w_t + v_t) - \nabla F_{S^{(i)}}(w_t^{(i)} + v_t^{(i)})\|^2}_{(I)} \\ &\quad - \underbrace{2\eta\langle \nabla F_S(w_t + v_t) - \nabla F_{S^{(i)}}(w_t^{(i)} + v_t^{(i)}), w_t - w_t^{(i)}\rangle}_{(II)}\end{aligned}$$

Treating the two terms (I),(II) separately, for (I) it holds by Lemma 8 that,

$$\begin{aligned}&\eta^2\|\nabla F_S(w_t + v_t) - \nabla F_{S^{(i)}}(w_t^{(i)} + v_t^{(i)})\|^2 \\ &\leq 2\eta^2\|\nabla F_{S^{(i)}}(w_t + v_t) - \nabla F_{S^{(i)}}(w_t^{(i)} + v_t^{(i)})\|^2 + \frac{2\eta^2}{n^2}\|\nabla f(w_t + v_t, z_i)\|^2 \\ &\leq 2\eta^2\|\nabla F_{S^{(i)}}(w_t + v_t) - \nabla F_{S^{(i)}}(w_t^{(i)} + v_t^{(i)})\|^2 + \frac{4\eta^2}{n^2}\|\nabla f(w_t + v_t, z_i\|^2 \\ &\leq 2\eta^2\|\nabla F_{S^{(i)}}(w_t + v_t) - \nabla F_{S^{(i)}}(w_t^{(i)} + v_t^{(i)})\|^2 + \frac{8\beta\eta^2}{n^2} f(w_t + v_t, z_i).\end{aligned}$$

For (II), it holds by two uses of Young's inequality that,

$$- 2\eta\langle \nabla F_S(w_t + v_t) - \nabla F_{S^{(i)}}(w_t^{(i)} + v_t^{(i)}), w_t - w_t^{(i)}\rangle$$

$$
\begin{aligned}
&= -2\eta\langle \nabla F_{S^{(i)}}(w_t+v_t) - \nabla F_{S^{(i)}}(w_t^{(i)}+v_t^{(i)}), w_t - w_t^{(i)}\rangle - \frac{2\eta}{n}\langle \nabla f(w_t+v_t,z_i), w_t - w_t^{(i)}\rangle \\
&= -2\eta\langle \nabla F_{S^{(i)}}(w_t+v_t) - \nabla F_{S^{(i)}}(w_t^{(i)}+v_t^{(i)}), w_t + v_t - w_t^{(i)} - v_t^{(i)}\rangle \\
&\quad - \frac{2\eta}{n}\langle \nabla f(w_t+v_t,z_i), w_t - w_t^{(i)}\rangle + 2\eta\langle \nabla F_{S^{(i)}}(w_t+v_t) - \nabla F_{S^{(i)}}(w_t^{(i)}+v_t^{(i)}), v_t - v_t^{(i)}\rangle \\
&\leq -\frac{2\eta}{\beta}\|\nabla F_{S^{(i)}}(w_t+v_t) - \nabla F_{S^{(i)}}(w_t^{(i)}+v_t^{(i)})\|^2 \\
&\quad + \frac{\eta}{\alpha n}\|w_t - w_t^{(i)}\|^2 + \frac{\eta\alpha}{n}\|\nabla f(w_t+v_t,z_i)\|^2 \\
&\quad + \frac{\eta}{\gamma}\|\nabla F_{S^{(i)}}(w_t+v_t) - \nabla F_{S^{(i)}}(w_t^{(i)}+v_t^{(i)})\| + \eta\gamma\|v_t - v_t^{(i)}\|^2
\end{aligned}
$$

By setting $\alpha = \eta T/n$ and using co-coercivity of-gradients of smooth functions, we get,

$$
\begin{aligned}
&-2\eta\langle \nabla F_S(w_t+v_t) - \nabla F_{S^{(i)}}(w_t^{(i)}+v_t^{(i)}), w_t - w_t^{(i)}\rangle \\
&\leq (\frac{\eta}{\gamma} - \frac{2\eta}{\beta})\|\nabla F_{S^{(i)}}(w_t+v_t) - \nabla F_{S^{(i)}}(w_t^{(i)}+v_t^{(i)})\|^2 + \frac{\eta}{\alpha n}\|w_t - w_t^{(i)}\|^2 \\
&\qquad\qquad\qquad\qquad + \frac{2\beta\alpha\eta}{n} f(w_t+v_t,z_i) + 4\eta\gamma r^2 \\
&\leq (\frac{\eta}{\gamma} - \frac{2\eta}{\beta})\|\nabla F_{S^{(i)}}(w_t+v_t) - \nabla F_{S^{(i)}}(w_t^{(i)}+v_t^{(i)})\|^2 + \frac{1}{T}\|w_t - w_t^{(i)}\|^2 \\
&\qquad\qquad\qquad\qquad + \frac{2\beta\eta^2 T}{n^2} f(w_t+v_t,z_i) + 4\eta\gamma r^2.
\end{aligned}
$$

Averaging over $i \in [n]$, plugging both in, and setting $\gamma = \beta$,$\eta \leq \frac{1}{2\beta}$

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|w_{t+1} - w_{t+1}^{(i)}\|^2 \\
&\leq \left(1 + \frac{1}{T}\right)\frac{1}{n}\sum_{i=1}^{n}\|w_t - w_t^{(i)}\|^2 + \frac{8\beta\eta^2(T+1)}{n^2}F_S(w_t+v_t) \\
&\qquad + 4\eta\gamma r^2 + (2\eta^2 - \frac{2\eta}{\beta} + \frac{\eta}{\gamma})\|\nabla F_{S^{(i)}}(w_t+v_t) - \nabla F_{S^{(i)}}(w_t^{(i)}+v_t^{(i)})\|^2 \\
&\leq \left(1 + \frac{1}{T}\right)\frac{1}{n}\sum_{i=1}^{n}\|w_t - w_t^{(i)}\|^2 + \frac{8\beta\eta^2(T+1)}{n^2}F_S(w_t+v_t) + 4\eta\beta r^2 \\
&\leq \frac{e^{\frac{1}{T}}}{n}\sum_{i=1}^{n}\|w_t - w_t^{(i)}\|^2 + \frac{8\beta\eta^2(T+1)}{n^2}F_S(w_t+v_t) + 4\eta\beta r^2.
\end{aligned}
$$

Now, unrolling the recursion, we get,

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\|w_{t+1} - w_{t+1}^{(i)}\|^2 &\leq \sum_{t=1}^{T} e^{\frac{T-t}{T}}\left(\frac{8\beta\eta^2(T+1)}{n^2}F_S(w_t+v_t) + 4\eta\beta r^2\right) \\
&\leq \frac{24\beta\eta^2(T+1)}{n^2}\sum_{t=1}^{T}F_S(w_t+v_t) + 12\eta\beta r^2 T.
\end{aligned}
$$

Using Lemma 1, we get for every t that,

$$\frac{1}{n}\sum_{i=1}^{n}\|w_{t+1}-w_{t+1}^{(i)}\|^2 \leq \frac{96\beta\eta^2 T}{n^2}\left(\frac{1}{\eta}+4\beta T\max(r-\rho,0)^2\right)+12\eta\beta r^2 T$$

$$=12\eta\beta r^2 T+\frac{96\beta\eta T}{n^2}+\frac{384\beta^2\eta^2T^2\max(r-\rho,0)^2}{n^2}.$$

By Jensen's inequality and the convexity of squared ℓ_2 norm, we get that,

$$\frac{1}{n}\sum_{i=1}^{n}\|\frac{1}{T}\sum_{i=1}^{T}w_t-\frac{1}{T}\sum_{i=1}^{T}w_t^{(i)}\|^2 \leq 24\eta\beta r^2 T+\frac{96\beta\eta T}{n^2}+\frac{768\beta^2\eta^2T^2\max(r-\rho,0)^2}{n^2}.$$

■

Proof of Theorem 7 By Theorem 4, we know that

$$F_S\left(\frac{1}{T}\sum_{t=1}^{T}w_t\right)\leq F_S^r\left(\frac{1}{T}\sum_{t=1}^{T}w_t\right)\leq\frac{\|w_1-w^\star\|^2}{\eta T}+4\beta\max\{r-\rho,0\}^2.$$

By Lemma 3, we know that, the algorithm is ℓ_2-on-average model ε-stable with

$$\varepsilon\leq 24\eta\beta r^2 T+\frac{96\beta\eta T}{n^2}+\frac{768\beta^2\eta^2T^2\max(r-\rho,0)^2}{n^2}.$$

By combining both equations with Lemma 2 we get the theorem. ■

Proof of Theorem 8 The proof is identical to the proof of Theorem 7 except for using Theorem 5 instead of Theorem 4. ■