Noise Stability Optimization for Finding Flat Minima: A Hessian-based Regularization Approach

Anonymous authors
Paper under double-blind review

Abstract

Consider a noise-injected function F of an input function f, perturbed by random noise U with mean zero. When the noise follows an isotropic Gaussian distribution, F is approximately f plus a penalty on the trace of the Hessian of f, scaled by noise variance. The natural approach of adding noise perturbation to weights before gradient computation in SGD shows limited improvement in fine-tuning pretrained networks on image classification datasets. We hypothesize that this is caused by the increased variance of the noise-injected gradient rather than the ineffectiveness of Hessian regularization. To address this issue, we propose a two-point noise injection algorithm, adding noise in both U and $-U$ and averaging over multiple perturbations to reduce gradient variance. A generalization bound related to the trace of the Hessian and the fine-tuned region's radius is shown to support this approach.

We present comprehensive experiments demonstrating that the two-point noise injection algorithm enhances generalization and Hessian regularization. The algorithm outperforms existing sharpness-reducing training methods, achieving up to a 1.8% increase in test accuracy for fine-tuning pretrained ResNets on six image classification datasets. It also results in a 17.7% reduction in the trace and a 12.8% reduction in the largest eigenvalue of the loss Hessian matrix. The Hessian regularization induced by noise injection is compatible with other popular regularization methods, such as weight decay and data augmentation, and their combination leads to improved empirical performance.

We present a detailed convergence analysis of the two-point noise injection algorithm, demonstrating precise rates on the norm of the gradient of the iterates. This analysis leverages techniques from the stochastic optimization literature, establishing a new link between these techniques and the analysis of sharpness-reducing methods.

1 Introduction

The loss landscape and geometry properties of neural networks have been widely studied (Keskar et al., 2017; Dinh et al., 2017), with research showing that flat loss surfaces can improve generalization (Hochreiter & Schmidhuber, 1997). Regularization schemes and methods like sharpness-aware minimization (SAM) (Foret et al., 2021) and stochastic weight averaging (Izmailov et al., 2018) have shown empirical success, particularly in low-sample regimes such as model fine-tuning (Wortsman et al., 2022a). Despite these advances, the theoretical properties of sharpness-reducing methods remain underexplored (Andriushchenko & Flammarion, 2022). Recent work (Wen et al., 2023) has characterized the implicit bias of SAM as a penalty on the largest eigenvalue of the loss Hessian matrix and noted that SAM's local minimizers oscillate around a local basin (Bartlett et al., 2023). Notably, SAM is based on a constrained min-max problem. In this paper, we study a min-avg problem, focusing on improving generalization for model fine-tuning: given an input function $f : \mathbb{R}^d \to \mathbb{R}$ such as the empirical risk and a d-dimensional distribution $\mathcal{P}$ with mean zero, we consider minimizing noise-perturbed functions $F(W) := \mathbb{E}_{U\sim\mathcal{P}}\left[f(W+U)\right]$.

The sensitivity or resilience of the input function f around its local neighborhood (measured by $F - f$) can influence algorithms to converge to wide minima (Nagarajan & Kolter, 2020). Research shows that optimizing prior and posterior distributions from data achieves generalization bounds that align with empirical

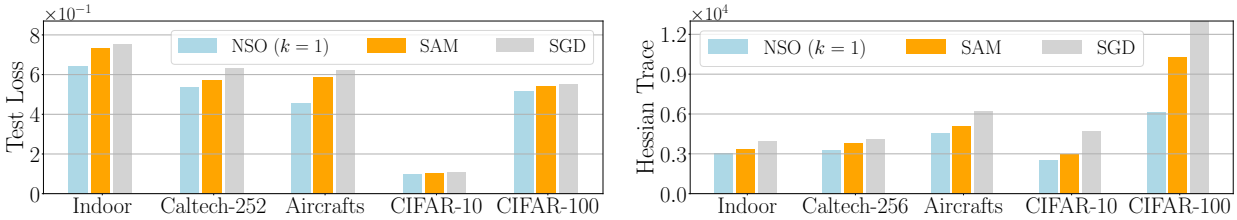


Figure 1: Illustration of the test loss (left) and the trace of the Hessian (right), measured at the last epoch of model fine-tuning. The results are run from a pretrained ResNet-34 network across five image classification tasks. We report the averaged results over five random seeds; Their standard deviations are reported in Section 3. Our proposed approach (NSO) can reduce both measures compared to SAM and SGD.

generalization gaps (Dziugaite & Roy, 2017). Additionally, the Hessian of multi-layer neural networks relates to noise sensitivity (Tsuzuku et al., 2020; Ju et al., 2022). However, both earlier (Hinton & Van Camp, 1993; An, 1996; Graves, 2011) and recent works (Orvieto et al., 2023) suggest that the regularization effect of noise injection in SGD is not always evident.

To replicate these findings in our setting, we compare the performance of standard SGD and noise-injected SGD for fine-tuning pretrained models on three classification datasets. It is observed that the noise injection does not offer clear benefits, even after testing several noise distributions. This may be due to the stochasticity of the noise injection, where the noise term's variance on the gradient can overshadow the second-order Hessian term, especially during a small number of epochs, such as transformer network fine-tuning.

To address this issue, we introduce two adjustments: i) adding a negative perturbation along $W - U$ to cancel out the first-order term while preserving the second-order term unaffected, and ii) sampling multiple perturbations per step to reduce gradient variance. Unlike SAM, where the effect from the signed gradient step is more nuanced (Wen et al., 2023; Andriushchenko et al., 2024), these adjustments provide an unbiased Hessian estimate.

This approach is justified with a generalization bound dependent on the trace of the Hessian in empirical risk minimization and the radius of the fine-tuned region, utilizing a linear PAC-Bayes bound (Catoni, 2007; McAllester, 2013; Dziugaite et al., 2021) and optimizing the noise variance.

Next, we provide comprehensive experiments to validate our approach compared to existing sharpness-reducing training methods. The key findings include:

- The algorithm improves generalization in an over-parameterized sensing problem, achieving the lowest test loss compared to both SGD and noise-injected SGD, all of which can find solutions with near-zero training loss.
- Across a wide range of image classification data sets, the algorithm finds neural networks with lower test loss and better-regularized Hessians than four sharpness-reducing methods. See Figure 1 for an illustration of the comparison.
- The regularization effect on the Hessian is compatible with other regularization methods such as weight decay, data augmentation, and distance-based regularization. Combining our approach with any of these methods leads to improved results.

Lastly, we analyze the convergence of our algorithm by establishing matching upper and lower bounds on the norm of the gradient of the iterates. The upper bound builds on classical results from the stochastic optimization literature (Ghadimi & Lan, 2013; Lan, 2020; Zhang, 2023), while carefully analyzing the gradient variance in our procedure. The lower bound leverages recent advances in showing query complexity lower bounds (Carmon et al., 2020; Drori & Shamir, 2020). Although the proof techniques are standard, their application to sharpness-reducing methods is novel. Our work implies that it is possible to design flat-minima optimizers with strong empirical performance and a clear analysis of convergence, which may benefit future research in this area.

In summary, the main results of this paper include:

1. Revisiting the generalization effect of noise injection for fine-tuning pretrained models and designing an algorithm that regularizes the trace of the Hessian.

2. Demonstrating that the proposed approach provides strong empirical performance compared to four sharpness-reducing training methods, and that the regularization induced by noise injection is compatible with other popular regularization methods.

3. Analyzing the convergence of the proposed algorithm using techniques from stochastic optimization, establishing a new connection to the analysis of sharpness-reducing methods.

1.1 Related Work

Methods for fine-tuning neural networks have garnered significant attention, as these methods are now widely used for adapting pretrained models (Hu et al., 2022; Wortsman et al., 2022b). Among these methods, sharpness-aware minimization (SAM) is particularly effective, motivated by a constrained min-max optimization problem, though it is computationally intractable (Daskalakis et al., 2021). Bartlett et al. (2023) found that for a convex quadratic function, SAM's stationary point oscillates locally according to the eigenvector corresponding to the largest eigenvalue, a behavior also observed in simulations. Additionally, ensemble methods have proven effective for improving the robustness of fine-tuning (Wortsman et al., 2022a). The generalization properties of fine-tuning in transformer networks are not well understood; for instance, examining the Hessian during the fine-tuning procedure could be a promising direction for future work.

The concept that injecting noise into neural networks can induce flatness in the found minima dates back to early research (Hinton & Van Camp, 1993; An, 1996). Graves (2011) develop a variational inference approach to test different priors and posteriors (e.g., Delta, Laplace, Uniform, Gaussian) on recurrent neural networks. Camuto et al. (2020) proposes a layer-wise regularization scheme motivated by adaptation patterns of weights through deeper layers. Bisla et al. (2022) conduct empirical studies on the connection between sharpness and generalization. Orvieto et al. (2023) analyze Taylor's expansion of the stochastic objective after noise injection, examining the induced regularization in various neural network training settings, and found that layer-wise perturbation can improve generalization and test accuracy. The PAC-Bayes analysis framework is related to noise injection, as it studies model generalization by postulating prior and posterior distributions on the hypothesis space (McAllester, 1999; Shawe-Taylor & Williamson, 1997).

The connection between Hessian and sharpness has also been studied through the Edge of Stability (Cohen et al., 2021), which is inverse to the operator norm of the Hessian matrix. Long & Bartlett (2023) identify the edge of stability regime for the SAM algorithm, highlighting differences from gradient descent. Additionally, Gaussian smoothing has been used to estimate gradients in zeroth-order optimization (Nesterov & Spokoiny, 2017). Besides, recent research has investigated the query complexity of finding stationary points of nonconvex functions (Carmon et al., 2020; Arjevani et al., 2023). These results provide a fine-grained characterization of the iteration complexity of iterative methods, under different orders of gradient oracles. There is a strong connection between sharpness and generalization, and we hope this work will inspire future research on the interplay between generalization and optimization.

Organization: The rest of this paper is organized as follows. In Section 2, we present the proposed approach. Then in Section 3, we describe the experiments conducted to validate the approach. In Section 4, we analyze the algorithm's convergence. In Section 5, we conclude this paper. In Appendix A and Appendix B, we show complete proofs of the theoretical statements.

2 Our Approach

In this section, we describe the regularization effect of our approach through the Hessian. We first state a proposition that describes the implicit bias of the population function $F(W)$ upon $f(W)$.

Proposition 2.1. *Suppose $f(W)$ is twice-differentiable. Let $\Sigma \in \mathbb{R}^{d\times d}$ be a positive semi-definite matrix. For a Gaussian distribution $\mathcal{P} = \mathcal{N}(0, \Sigma)$ and a random sample $U \in \mathbb{R}^d$ drawn from $\mathcal{P}$, the following holds with high probability:*

$$F(W) = \mathop{\mathbb{E}}_{U\sim\mathcal{P}}\left[\frac{1}{2}(f(W+U) + f(W-U))\right] = f(W) + \frac{1}{2}\langle \Sigma, \nabla^2 f(W)\rangle + O(\|\Sigma\|_2^{\frac{3}{2}}). \tag{1}$$

Recall that $\mathcal{P}$ is a symmetric distribution. We use Taylor's expansion on both $f(W+U)$ and $f(W-U)$:

$$f(W+U) = f(W) + \langle U, \nabla f(W)\rangle + \frac{1}{2}U^\top \nabla^2 f(W) U + O(\|\Sigma\|_2^{\frac{3}{2}}),$$
$$f(W-U) = f(W) - \langle U, \nabla f(W)\rangle + \frac{1}{2}U^\top \nabla^2 f(W) U + O(\|\Sigma\|_2^{\frac{3}{2}}).$$

By definition, $\mathbb{E}[U] = 0$, and $\mathbb{E}\left[UU^\top\right] = \Sigma$. Thus, by taking the average of the above two equations, one can get equation (1).

As a remark, we point out that the case without the injection in the negative direction of U, has been noted in prior work (Orvieto et al., 2023). As we shall show next, adding the negative perturbation step is crucial for the algorithm to succeed. Concretely, we will study the behavior of noise injection for fine-tuning pre-trained neural networks, as overfitting is quite common in this setting (Wortsman et al., 2022a). Hence, developing methods to improve the generalization of fine-tuning would be crucial.

Experimental Setup: We consider fine-tuning a pre-trained ResNet-34 on image classification data sets, including aircraft recognition (Aircrafts) (Maji et al., 2013), indoor scene recognition (Caltech-256) (Griffin et al., 2007), and medical image classification (retina images for diabetic retinopathy classification) (Pachade et al., 2021). We will compare i) vanilla SGD, and ii) weight-perturbed SGD (or WP-SGD in short), where we sample a perturbation vector from $\mathcal{P}$ and add it to the model weights in each iteration before computing the gradient. For WP-SGD, we will sample the perturbation from an isotropic Gaussian distribution. Then, we will set the standard deviation of the Gaussian based on validation performance, chosen between $0.008, 0.01, 0.012$.

Summary of Findings: We report our findings in Table 1, listed as follows:

- We observe that the performance gap between SGD and WP-SGD is within 0.5%, accounting for the standard deviations of the individual runs.
- Varying the type of noise distribution does not change the result. In particular, we test alternative choices of $\mathcal{P}$ with Laplace distribution, uniform distribution, and Binomial distribution. Similar to the Gaussian, we set their standard deviations between $0.008, 0.01, 0.012$ using a validation set.
- Using the Laplace or Uniform distribution achieves a performance comparable to Gaussian. However, WP-SGD struggles to converge using the Binomial distribution, resulting in significantly lower training and test results.

Based on the above empirical findings, we now present our approach, which involves two components:

- *Two-point noise injection:* During the noise injection, we add the perturbation from both the positive direction and the negative direction. This is shown in Line 4.
- *Averaging multiple perturbations to stabilize the gradient:* To stabilize the stochasticity of the noise injection, we average over multiple noise injections. This is described in Line 6.

The full procedure is summarized in Algorithm 1 below. As a remark, a two-point estimator has been used to estimate the gradient in the setting of zero-order convex optimization (Duchi et al., 2015). The difference is that we use the two-point average of $W+U$ and $W-U$ to cancel out the first-order gradient term.

Table 1: Comparing weight perturbed SGD (WP-SGD) to SGD, across four types of perturbation distributions, measured over three image classification data sets. The results and their standard deviations are averaged over five independent seeds.

		Aircrafts		Indoor		Retina	
	$\mathcal{P}$	**Train Acc.**	**Test Acc.**	**Train Acc.**	**Test Acc.**	**Train Acc.**	**Test Acc.**
SGD	None	100.0% ± 0.0	59.8% ± 0.7	100.0% ± 0.0	76.0% ± 0.4	100.0% ± 0.0	61.7% ± 0.8
WP-SGD	Gaussian	98.4% ± 0.2	60.4% ± 0.1	99.0% ± 0.3	76.3% ± 0.0	100.0% ± 0.0	62.3% ± 0.5
WP-SGD	Laplace	98.3% ± 0.1	60.3% ± 0.3	98.9% ± 0.1	76.4% ± 0.3	100.0% ± 0.0	62.0% ± 0.1
WP-SGD	Uniform	98.6% ± 0.3	60.3% ± 0.5	98.6% ± 0.3	76.6% ± 0.1	100.0% ± 0.0	62.3% ± 0.0
WP-SGD	Binomial	19.6% ± 0.1	11.3% ± 0.1	18.2% ± 0.9	10.7% ± 0.1	58.1% ± 0.1	57.1% ± 0.0

Algorithm 1 Noise stability optimization (NSO) for improving generalization of neural nets

Input: Initialization $W_0 \in \mathbb{R}^d$, a function $f : \mathbb{R}^d \to \mathbb{R}$
Require: An estimator $g : \mathbb{R}^d \to \mathbb{R}^d$ that for any W, returns $g(W)$ s.t. $\mathbb{E}\left[g(W)\right] = \nabla f(W)$
Parameters: # perturbations k, # epochs T, step sizes $\eta_0, \ldots, \eta_{T-1}$

1: **for** $i = 0, 1, \ldots, T-1$ **do**
2: **for** $j = 0, 1, \ldots, k-1$ **do**
3: Sample $U_i^{(j)}$ independently from $\mathcal{P}$
4: Let $G_i^{(j)} = g\big(W_i + U_i^{(j)}\big) + g\big(W_i - U_i^{(j)}\big)$
5: **end for**
6: Update iterates according to $W_{i+1} = W_i - \frac{\eta_i}{2k} \sum_{j=1}^{k} G_i^{(j)}$
7: **end for**

2.1 A Generalization Bound Using Trace of the Hessian

Next, we present a PAC-Bayes bound which shows that the trace of the Hessian serves as an upper bound on the generalization gap. As a remark, the trace norm has been studied by earlier work in the setting of matrix recovery (Srebro & Shraibman, 2005). Its use as a generalization measure for fine-tuning has not been studied before, up to our knowledge.

Concretely, in the fine-tuning setting, we have a pretrained model, which can be viewed as the prior in PAC-Bayes analysis. Once we have learned a hypothesis, it can be viewed as the posterior. Let $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$ be an unknown data distribution, supported on the feature space $\mathcal{X}$ and the label space $\mathcal{Y}$. Given n random samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ drawn from $\mathcal{D}$, the empirical loss (measured by loss function ℓ) applied to a model f_W (with $W \in \mathbb{R}^p$) is:

$$\hat{L}(W) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_W(x_i), y_i).$$

The population loss is $L(W) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(f_W(x), y)\right]$. It is sufficient to think that the empirical loss is less than the population loss, and the goal is to bound the gap from above (Shalev-Shwartz & Ben-David, 2014).

Let W be any learned hypothesis within the hypothesis space, denoted as $\mathcal{H}$. The generalization bound will apply uniformly to W within the hypothesis space, assuming that this space, centered at the pretrained initialization, has a bounded radius of $r > 0$. We state the result as follows.

Theorem 2.2. *Assume that the loss function is bounded between* 0 *and* C*, for a fixed constant* C*. Suppose that* $\ell(f_W(\cdot), \cdot)$ *is twice-differentiable in* W *and the Hessian matrix* $\nabla^2_W[\ell(f_W(\cdot), \cdot)]$ *is Lipschitz continuous within the hypothesis space. With probability at least* $1 - \delta$ *for any* $\delta > 0$*, the following must hold, for any* ϵ *close to zero:*

$$L(W) \leq (1+\epsilon)\hat{L}(W) + (1+\epsilon)\sqrt{\frac{C\alpha r^2}{n}} + O\Big(n^{-\frac{3}{4}} \log(\delta^{-1})\Big). \tag{2}$$

where $\alpha := \max_{W\in\mathcal{H}} \max_{(x,y)\sim\mathcal{D}} \operatorname{Tr}\left[\nabla^2 \ell(f_W(x), y)\right]$ *is the trace norm of the hypothesis space taken over the data distribution* $\mathcal{D}$.

Proof Sketch: We provide a high-level illustration of the ideas behind Theorem 2.2 without belaboring too much on the technical details. Let $\mathcal{Q}$ denote the *posterior* distribution. Specifically, we consider $\mathcal{Q}$ as being centered at the learned hypothesis W (which could be anywhere within the hypothesis space), given by a Gaussian distribution $\mathcal{N}(W, \sigma^2 \operatorname{Id}_p)$, where Id_p denotes the p by p identify matrix. Given a sample $U \sim \mathcal{N}(0, \sigma^2 \operatorname{Id}_p)$, let the perturbed loss $\ell_{\mathcal{Q}}(f_W(x), y)$ be given by $\mathbb{E}_U\left[\ell(f_{W+U}(x), y)\right]$. Then, let $\hat{L}_{\mathcal{Q}}(W)$ be the averaged value of $\ell_{\mathcal{Q}}(f_W(\cdot), \cdot)$, taken over the n empirical samples. Likewise, let $L_{\mathcal{Q}}(W)$ be the population average of $\ell_{\mathcal{Q}}(f_W(\cdot), \cdot)$.

Having introduced the notations, one starts with the PAC-Bayes bound (Catoni, 2007; McAllester, 2013; Alquier, 2021) (see Theorem A.1 for reference), stated as follows:

$$L_{\mathcal{Q}}(W) \leq \frac{1}{\beta}\hat{L}_{\mathcal{Q}}(W) + \frac{C(KL(\mathcal{Q}||\mathcal{P}) + \log(\delta^{-1}))}{2\beta(1-\beta)n}, \tag{3}$$

where β is a parameter chosen between $(0, 1)$, and $\mathcal{P}$ is a *prior* distribution. For the fine-tuning setting, $\mathcal{P}$ can be viewed as centered at the pretrained initialization, with variance $\sigma^2 \operatorname{Id}_p$ similar to $\mathcal{Q}$.

Next, by Taylor's expansion of $\ell_{\mathcal{Q}}$ like Proposition 2.1 (see Lemma A.4 for the full result), we show that:

$$L_{\mathcal{Q}}(W) = L(W) + \frac{\sigma^2}{2} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\operatorname{Tr}\left[\nabla^2 \ell(f_W(x), y)\right]\right] + O(\sigma^3), \text{ and}$$

$$\hat{L}_{\mathcal{Q}}(W) = \hat{L}(W) + \frac{\sigma^2}{2n}\sum_{i=1}^{n} \operatorname{Tr}\left[\nabla^2 \ell(f_W(x_i), y_i)\right] + O(\sigma^3).$$

Since we have required the Hessian to be Lipschitz continuous, we bound the gap between the above two using uniform convergence on Lipschitz functions (see Lemma A.5 for the result). By plugging in the above two results back to the PAC-Bayes bound of equation (3), and making up the difference between $1/\beta$ and 1 between the left and right sides by α, we get:

$$L(W) \leq \frac{1}{\beta}\hat{L}(W) + \frac{\sigma^2(1-\beta)\alpha}{2\beta} + \frac{Cr^2/2\sigma^2}{2\beta(1-\beta)n} + O\left(\sigma^3 + \frac{\sigma^2\sqrt{p}}{\sqrt{n}} + \frac{\log(\delta^{-1})}{n}\right).$$

In particular, the above uses the fact that the hypothesis space is uniformly bounded in a ball of radius r, and the derivation of the KL divergence can be found in Proposition A.2. By choosing σ^2 and β to minimize the above bound, we thus obtain the result of equation (2). This summarizes the high-level proof idea. The complete proof can be found in Appendix A.1.

2.1.1 An In-depth Look of the Hessian in Matrix Sensing

Before proceeding, let us give an example to better understand the regularization effect of the Hessian. We consider the matrix sensing problem, whose generalization properties are particularly well-understood in the nonconvex factorization setting (Li et al., 2018). Let there be an unknown, rank-r positive semi-definite matrix $X^\star = U^\star U^{\star\top} \in \mathbb{R}^{d\times d}$. The input consists of a list of d by d Gaussian measurement matrix $A_1, A_2, \ldots, A_n$. The labels are given by $y_i = \langle A_i, X^\star\rangle$, for every $i = 1, 2, \ldots n$. The empirical loss is

$$\hat{L}(W) = \frac{1}{2n}\sum_{i=1}^{n}\left(\langle A_i, WW^\top\rangle - y_i\right)^2, \text{ where } W \in \mathbb{R}^{d\times d}. \tag{4}$$

When the loss reaches near zero (which implies the gradient also reaches near zero), it is known that multiple local minimum solutions exist (Li et al., 2018), and the Hessian becomes

$$\frac{1}{n}\sum_{i=1}^{n}\|A_i W\|_F^2 \approx d\,\|W\|_F^2 = d\left\|WW^\top\right\|_\star.$$

By prior results (Recht et al., 2010), among all $X = WW^\top$ such that $\hat{L}(W) = 0$, $X^\star$ has the lowest nuclear norm. Thus, the regularization placed on $\hat{L}(W)$ is similar to nuclear norm regularization under interpolation. We formalize this and state the proof below for completeness.

Proposition 2.3. *In the setting above, for any W that satisfies $\hat{L}(W) = 0$, the following must hold with high probability:*

$$\mathrm{Tr}\left[\nabla^2[\hat{L}(U^\star)]\right] \leq \mathrm{Tr}\left[\nabla^2[\hat{L}(W)]\right] + O(n^{-\frac{1}{2}}). \tag{5}$$

A similar statement holds if the trace operator is replaced by the largest eigenvalue of the Hessian in equation (5). To see this, we look at the quadratic form of the Hessian in order to find the maximum eigenvalue. Let u be a d^2 dimension vector with length equal to one, $\|u\| = 1$. One can derive that:

$$\lambda_1(\nabla^2\hat{L}(W)) = \max_{u\in\mathbb{R}^{d^2}:\|u\|=1} u^\top\nabla^2\hat{L}(W)u = \max_{u\in\mathbb{R}^{d^2}:\|u\|=1} \frac{1}{n}\sum_{i=1}^{n}\langle A_iW, u\rangle^2 \geq \frac{1}{d^2n}\sum_{i=1}^{n}\|A_iW\|_F^2 .$$

The last step is by setting $u = d^{-1}\mathbf{1}_{d^2}$, whose length is equal to one. The detailed proof of Proposition 2.3 and derivations for the above step are deferred in Appendix A.2.

2.1.2 Discussions

In the case that f is a strongly convex function, the lowest eigenvalue of the Hessian is bounded from below. Once the algorithm reaches the global minimizer, our result from Theorem 2 can be used to provide a generalization bound based on the trace of the Hessian. Notice that the noise injection will add some bias to this minimizer, leading to a sub-optimal empirical loss. To remedy this issue, one can place the regularization of the Hessian as a constraint, similar to how ℓ_2-regularization can be implemented as a constraint.

2.2 Empirical Measurements of the Hessian

Next, we provide several empirical examples to validate the theoretical bounds. Following the experimental setup described earlier, we fine-tune several pretrained models on one downstream task. We test on three different modalities of data, including images, texts, and graphs. After fine-tuning, we set the fine-tuned model weight at the last epoch as W, for taking all the measurements. We summarize the empirical findings below, leaving experimental details to Appendix C. First, we show that Taylor's expansion of the noise injection is numerically accurate. We add perturbations to model weights by injecting isotropic Gaussian noise. We then compute the perturbed loss minus the original loss value, averaged over 100 independent runs, and we measure the trace of the Hessian as the average over the training dataset.

- In Table 2, we find that the trace of the Hessian provides an accurate approximation to the gap between $\ell_\mathcal{Q}$ (recall that $\ell_\mathcal{Q}(f_W(x), y) = \mathbb{E}_U\left[\ell(f_{W+U}(x), y)\right]$) and ℓ. After fine-tuning, we add random noise injections to the fine-tuned model weight. We do this for 100 times, and we measure the perturbed loss $\ell_\mathcal{Q}$ again on the training set. We take the gap between $\ell_\mathcal{Q}$ and ℓ and report that along with the magnitude of σ in the table. We also compute the trace of the Hessian using Hessian-vector product computation libraries. Our measurements show that the error between the actual gap and the Hessian approximation is within 3%. As a remark, the range of σ^2 differs across architectures because of the differing scales of their weights.

- We compare the measurements between SGD and NSO in Figure 2. Curiously, we find that as the test loss goes down, the trace of the Hessian also goes down. While both SGD and NSO reduce the trace of the Hessian, our approach indeed penalizes the Hessian more significantly than SGD.

- Compared with SGD, the generalization gap of the fine-tuned model also lowers by over 20%. The test loss of the fine-tuned model using our approach is also lower than SGD.

Table 2: We find that the trace of the Hessian provides an accurate approximation to the gap between $\ell_{\mathcal{Q}}$ (recall that $\ell_{\mathcal{Q}}$ is the perturbed loss) and ℓ. In particular, the measurements are taken on the fine-tuned model weight W at the last epoch.

Multi-Layer Perceptron (MNIST)			BERT Base (MRPC)			Graph ConvNets (COLLAB)		
$\boldsymbol{\sigma}$	Gap	$\frac{\sigma^2}{2}\mathbf{Tr}\left[\nabla^2 f(W)\right]$	$\boldsymbol{\sigma}$	**Gap**	$\frac{\sigma^2}{2}\mathbf{Tr}\left[\nabla^2 f(W)\right]$	$\boldsymbol{\sigma}$	**Gap**	$\frac{\sigma^2}{2}\mathbf{Tr}\left[\nabla^2 f(W)\right]$
0.020	**0.0122** $\pm$ 0.0027	**0.0096**	0.0070	**0.0083** $\pm$ 0.0031	**0.0095**	0.040	**0.0297** $\pm$ 0.0097	**0.0278**
0.021	**0.0124** $\pm$ 0.0026	**0.0106**	0.0071	**0.0088** $\pm$ 0.0031	**0.0098**	0.041	**0.0266** $\pm$ 0.0141	**0.0292**
0.022	**0.0137** $\pm$ 0.0042	**0.0117**	0.0072	**0.0093** $\pm$ 0.0032	**0.0101**	0.042	**0.0363** $\pm$ 0.0086	**0.0306**
0.023	**0.0142** $\pm$ 0.0049	**0.0128**	0.0073	**0.0098** $\pm$ 0.0034	**0.0103**	0.043	**0.0243** $\pm$ 0.0109	**0.0321**
0.024	**0.0152** $\pm$ 0.0046	**0.0139**	0.0074	**0.0104** $\pm$ 0.0035	**0.0106**	0.044	**0.0287** $\pm$ 0.0111	**0.0336**
0.025	**0.0175** $\pm$ 0.0047	**0.0151**	0.0075	**0.0110** $\pm$ 0.0036	**0.0109**	0.045	**0.0298** $\pm$ 0.0092	**0.0351**
0.026	**0.0182** $\pm$ 0.0038	**0.0163**	0.0076	**0.0117** $\pm$ 0.0038	**0.0112**	0.046	**0.0414** $\pm$ 0.0105	**0.0367**
0.027	**0.0209** $\pm$ 0.0035	**0.0176**	0.0077	**0.0124** $\pm$ 0.0040	**0.0115**	0.047	**0.0313** $\pm$ 0.0109	**0.0383**
0.028	**0.0215** $\pm$ 0.0049	**0.0189**	0.0078	**0.0131** $\pm$ 0.0042	**0.0118**	0.048	**0.0455** $\pm$ 0.0089	**0.0400**
0.029	**0.0244** $\pm$ 0.0075	**0.0203**	0.0079	**0.0139** $\pm$ 0.0044	**0.0121**	0.049	**0.0449** $\pm$ 0.0160	**0.0417**
0.030	**0.0258** $\pm$ 0.0059	**0.0218**	0.0080	**0.0147** $\pm$ 0.0047	**0.0124**	0.050	**0.0482** $\pm$ 0.0100	**0.0434**
RSS	**2.74%**			**1.03%**			**2.16%**	

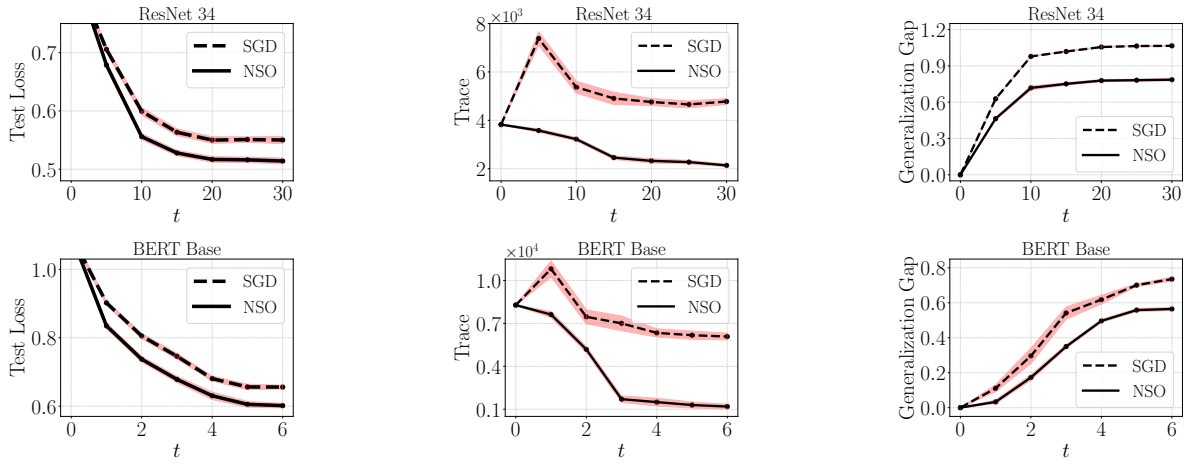


Figure 2: Comparison between SGD and NSO, for fine-tuning ResNet-34 and BERT-Base, on an image and a text classification data set, respectively. We report the test loss, the trace of the Hessian, and the generalization gap, for W taken at the last epoch. For NSO, we sample random perturbations using isotropic Gaussian distribution with standard deviation $\sigma = 0.01$ for both settings.

3 Experimental Results

We now turn to the empirical validation of our proposed algorithm. Through both simulations in the matrix sensing problem and experiments in fine-tuning neural networks, we show that our algorithm can indeed improve generalization, and this improvement can be explained by the regularization of the Hessian.

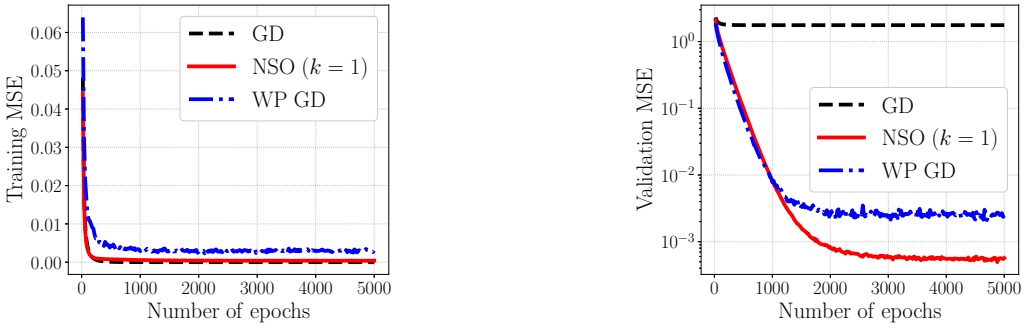
- Across various image classification data sets, NSO can outperform four previous sharpness-reducing methods by up to **1.8**%. We control the amount of computation in the experiments to allow for a fair comparison. We justify each step of the algorithm design, through ablation analysis.
- We notice that NSO regularizes the Hessian of the loss surface much more significantly, by noting reductions in the trace (and the largest eigenvalue) of the loss Hessian by **17.7**% (and **12.8**%), respectively.
- Our method is compatible with existing regularization techniques, including weight decay, distance-based regularization, and data augmentation, as combining these techniques leads to even greater improvement in both the Hessian regularization and the test performance.

3.1 Simulation

We conduct a numerical simulation in the matrix sensing problem. We generate a low-rank matrix $U^\star \in \mathbb{R}^{d\times r}$ from the isotropic Gaussian. We set $d = 100$ and $r = 5$. Then, we test three algorithms: gradient descent (GD), weight-perturbed gradient descent (WP-GD), and Algorithm 1 (NSO). We use an initialization $U_0 \in \mathbb{R}^{d\times d}$ where each matrix entry is sampled independently from $\mathcal{N}(0, 1)$ (the standard Gaussian).

Recall that WP-GD and NSO require setting σ. We choose σ between $0.001, 0.002, 0.004, 0.008, 0.0016$. NSO additionally requires setting the number of sampled perturbations k. We set $k = 1$ for faster computation.

Our findings are illustrated in Figure 3. We can see that all three algorithms can reduce the training MSE to near zero, as shown in Figure 3a. Regarding the validation loss, GD suffers from overfitting the training data, while both WP GD and NSO can generalize to the validation samples. Moreover, NSO manages to reduce this validation loss further.



(a) Training loss

(b) Validation loss

Figure 3: Comparing the training and validation losses between GD, NSO, and WP-GD.

3.2 Comparison with Sharpness Reducing Training Methods

We now compare Algorithm 1 with four sharpness-reducing training methods, including Sharpness-Aware Minimization (SAM) (Foret et al., 2021), Adaptive SAM (ASAM) (Kwon et al., 2021), Random SAM (RSAM) (Liu et al., 2022), and Bayesian SAM (BSAM) (Möllenhoff & Khan, 2023). During comparison, we control for the same amount of computation, by setting the number of sampled injections $k = 1$. Thus, all of these methods will use twice the cost of SGD in the end. For NSO, we sample perturbation from an isotropic Gaussian distribution and tune σ between $0.008, 0.01$, and 0.012 using a validation split. For SAM, we tune the ℓ_2 norm of the perturbation between $0.01, 0.02$, and 0.05. Since each other training method involves its own set of hyper-parameters, we make sure they are carefully selected. The details are tedious; See Appendix C for the range of values used for each hyper-parameter. To calibrate these results, we include both SGD and Label Smoothing (LS), as they are both widely used in practice.

We report the overall comparison in Table 3. In a nutshell, NSO performs competitively with all the baseline variants. Across these six data sets, NSO can achieve up to **1.8**% accuracy gain, with an average test accuracy improvement of **0.9**%, relative to the best-performing baselines. The results are aggregated over five independent runs, suggesting that our findings are statistically significant.

Ablation Analysis: Next, we conduct ablation studies of two components in NSO, i.e., using negative perturbations and sampling multiple perturbations in each iteration, showing both are essential.

Comparing using vs. not using negative perturbations: Recall that our algorithm uses negative perturbations to zero out the first-order order in Taylor's expansion of $F(W)$, leading to a better estimation of $\nabla F(W)$. We validate this by comparing the performance between using and not using the negative perturbation. To ensure that both use the same amount of computation, we sample two independent perturbations when not

Table 3: Comparison between NSO, SGD, Label Smoothing (LS), SAM, Adaptive SAM, Random-SAM, and Bayesian SAM, on six image classification data sets, by fine-tuning a pre-trained ResNet-34 neural network using each method. In this table, we report the test accuracy, the trace of the Hessian (for model weight found in the last epoch of each training algorithm), and also the largest eigenvalue of the Hessian. For the latter two measures, lower values indicate wider loss surfaces. In all test cases, we report the averaged result over five random seeds, and the standard deviation across these five runs. The results indicate that NSO outperforms the baselines in terms of the three metrics.

		CIFAR-10	**CIFAR-100**	**Aircrafts**	**Caltech-256**	**Indoor**	**Retina**
Basic Stats	Train	45,000	45,000	3,334	7,680	4,824	1,396
	Val.	5,000	5,000	3,333	5,120	536	248
	Test	10,000	10,000	3,333	5,120	1,340	250
	Classes	10	100	100	256	67	5
Test Acc. (↑)	SGD	95.5% ± 0.1	82.3% ± 0.1	59.8% ± 0.7	75.5% ± 0.1	76.0% ± 0.4	61.7% ± 0.8
	LS	96.7% ± 0.1	83.8% ± 0.1	58.5% ± 0.2	76.0% ± 0.2	75.9% ± 0.3	63.6% ± 0.7
	SAM	96.6% ± 0.4	83.5% ± 0.1	61.5% ± 0.8	76.3% ± 0.1	76.6% ± 0.5	64.4% ± 0.6
	ASAM	96.7% ± 0.1	83.8% ± 0.1	62.0% ± 0.6	76.7% ± 0.2	76.7% ± 0.3	64.8% ± 0.3
	RSAM	96.4% ± 0.1	83.7% ± 0.2	60.5% ± 0.5	75.8% ± 0.2	76.1% ± 0.7	65.4% ± 0.3
	BSAM	96.4% ± 0.0	83.5% ± 0.2	60.5% ± 0.5	76.3% ± 0.3	75.7% ± 0.7	64.9% ± 0.0
	NSO	**97.1%** ± 0.2	**84.3%** ± 0.2	**62.3%** ± 0.3	**77.4%** ± 0.3	**77.4%** ± 0.5	**66.6%** ± 0.7
Trace $\times 10^3$ (↓)	SGD	4.7 ± 0.0	14.4 ± 0.3	6.2 ± 0.0	4.1 ± 0.0	4.1 ± 0.0	30.4 ± 0.2
	LS	2.9 ± 0.0	11.3 ± 0.4	6.3 ± 0.0	3.8 ± 0.0	4.2 ± 0.0	19.2 ± 0.1
	SAM	2.8 ± 0.0	10.2 ± 0.4	5.0 ± 0.0	3.8 ± 0.0	3.8 ± 0.0	16.4 ± 0.2
	ASAM	2.8 ± 0.0	10.5 ± 0.3	5.0 ± 0.0	3.8 ± 0.0	3.1 ± 0.0	14.7 ± 0.1
	RSAM	2.7 ± 0.0	10.3 ± 0.5	5.5 ± 0.2	3.5 ± 0.0	4.1 ± 0.0	19.9 ± 0.5
	BSAM	3.0 ± 0.1	10.3 ± 0.5	5.6 ± 0.1	3.9 ± 0.0	3.5 ± 0.0	18.2 ± 0.3
	NSO	**2.2** ± 0.0	**5.9** ± 0.0	**4.2** ± 0.0	**3.3** ± 0.0	**3.0** ± 0.0	**11.6** ± 0.0
$\boldsymbol{\lambda_1}$ $\times 10^3$ (↓)	SGD	1.5 ± 0.0	4.9 ± 0.1	1.2 ± 0.0	1.1 ± 0.0	1.2 ± 0.1	9.0 ± 0.1
	LS	1.4 ± 0.0	3.5 ± 0.1	1.3 ± 0.1	1.0 ± 0.1	0.9 ± 0.1	4.9 ± 0.0
	SAM	1.4 ± 0.0	2.8 ± 0.1	0.9 ± 0.1	1.0 ± 0.0	1.0 ± 0.1	4.2 ± 0.0
	ASAM	1.4 ± 0.1	2.8 ± 0.1	0.6 ± 0.1	0.8 ± 0.0	**0.7** ± 0.1	4.2 ± 0.0
	RSAM	1.4 ± 0.1	3.0 ± 0.1	0.9 ± 0.1	0.8 ± 0.1	1.0 ± 0.0	5.0 ± 0.0
	BSAM	1.4 ± 0.0	3.0 ± 0.1	1.0 ± 0.1	0.9 ± 0.0	1.0 ± 0.1	4.3 ± 0.2
	NSO	**1.1** ± 0.1	**2.2** ± 0.1	**0.5** ± 0.1	**0.6** ± 0.0	**0.7** ± 0.1	**3.9** ± 0.0

using negative perturbations. We find that using negative perturbations achieves a **1.8**% improvement in test accuracy on average over the one without negative perturbations.

Varying the number of noise injections per iteration: Furthermore, increasing the number of perturbations k reduces the variance of the estimated $\nabla F(W)$. Thus, we consider increasing k in NSO and compare that with WP SGD with comparable computation. We find that using $k = 2$ perturbations improves the test accuracy by **1.2**% on average compared to $k = 1$. However, increasing k over 3 brings no obvious improvement (but adds more compute cost).

3.3 Compatibility with Existing Regularization Methods

Table 3 also shows the regularization effect of each training method on the Hessian. We compute the trace and the λ_1 of the loss Hessian matrix using power iteration implemented by Hessian vector product operations in PyTorch. Notably, in the middle and lower tables, where lower sharpness means better, NSO can significantly reduce them compared to the baselines, averaging **17.7**% (on trace) and **12.8**% (on λ_1).

The regularization on the Hessian can serve as a complement to existing regularization methods, including weight decay, distance-based regularization, and data augmentation. We combine NSO with these methods in the same experiment setup to validate this. For distance-based regularization, we penalize the ℓ_2 distance

from the fine-tuned model to the pre-trained initialization. For data augmentation, we use a popular scheme that sequentially applies random horizontal flipping and random cropping to each training image.

The results are shown in Figure 4. We confirm that combining our algorithm with each regularization method further reduces the trace of the loss Hessian matrix by **13.6**% on average. Quite strikingly, this also leads to **16.3**% lower test loss of the fine-tuned neural network on average, suggesting that our method can be used on top of these already existing regularization methods.

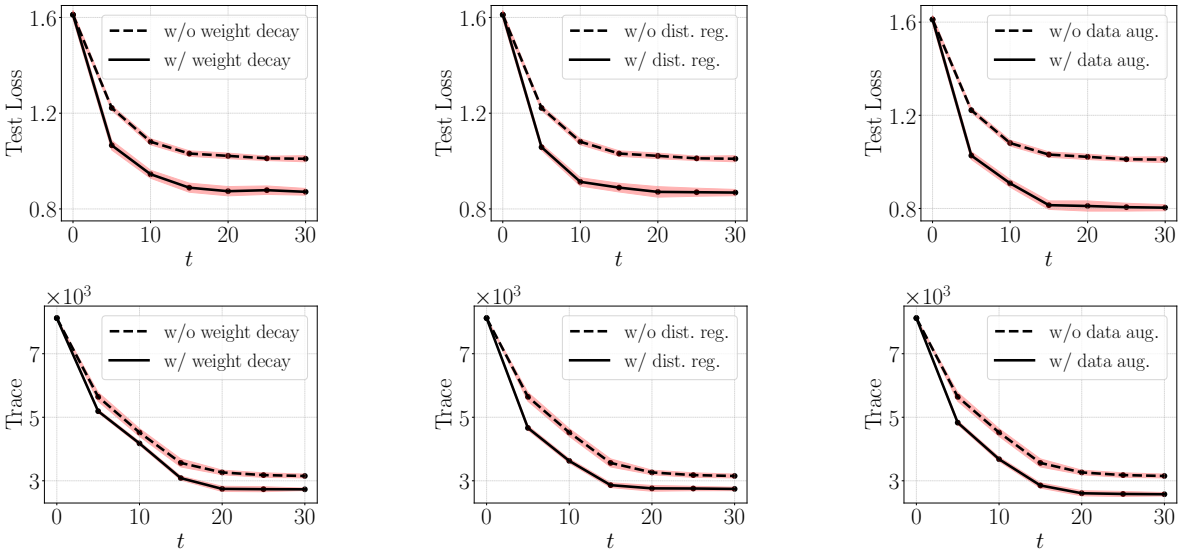


Figure 4: The Hessian regularization can be used in compatible with weight decay, ℓ_2 distance-based regularization, and data augmentation. We illustrate this for fine-tuning a pre-trained ResNet-34 neural network on an image classification data set. Combining each regularization method with ours generally leads to lower test losses and lowers the trace of the Hessian of the loss surface. Note that the shaded area indicates the deviation across five independent runs, suggesting the statistical significance of these findings.

4 Convergence Analysis

We now study the convergence of Algorithm 1. Recall that our algorithm minimizes $f(W)$ plus a regularization term on the trace of Hessian. As is typical with regularization, the penalty is usually small relative to the loss value. Thus, our goal is to find a stationary point of $F(W)$ instead of $f(W)$ because otherwise, we would not have the desired Hessian regularization. We state the convergence to an ϵ-approximate stationary point such that $\|\nabla F(W)\| \leq \epsilon$, for any small values of $\epsilon > 0$. The analysis builds on standard assumptions from the literature (Ghadimi & Lan, 2013; Duchi et al., 2015; Lan, 2020; Zhang, 2023).

Assumption 4.1. *Given a random seed z, let $g_z : \mathbb{R}^d \to \mathbb{R}^d$ be a continuous function that gives an unbiased estimate of the gradient: $\mathbb{E}_z\left[g_z(W)\right] = \nabla f(W)$, for any $W \in \mathbb{R}^d$. Additionally, the variance is bounded in the sense that $\mathbb{E}_z\left[\|g_z(W) - \nabla f(W)\|^2\right] \leq \sigma^2$.*

Assumption 4.2. *Let C, D be fixed, positive constants. Let $W_0 \in \mathbb{R}^d$ denote the initialization. We require that $F(W_0) - \min_{W \in \mathbb{R}^d} F(W) \leq D^2$. Let $\nabla f(W)$ denote the gradient of $f(W)$. For any $W_1 \in \mathbb{R}^d$ and $W_2 \in \mathbb{R}^d$, we have $\|\nabla f(W_2) - \nabla f(W_1)\| \leq C\|W_2 - W_1\|$. A corollary is that $\nabla F(W)$ is also C-Lipschitz.*

We now state an upper bound on the norm of the gradient of the returned solution.

Theorem 4.3. *Given Assumptions 4.1 and 4.2, let $\mathcal{P}$ be a distribution that is symmetric at zero. There exists a fixed learning rate $\eta < C^{-1}$ such that if we run Algorithm 1 with $\eta_i = \eta$ for all i, arbitrary number of perturbations k, for T steps, the algorithm returns W_t, where t is a random integer between $1, 2, \ldots, T$,*

such that in expectation over the randomness of W_t:

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \leq \sqrt{\frac{2CD^2(\sigma^2 + C^2 H(\mathcal{P}))}{kT}} + \frac{2CD^2}{T}, \tag{6}$$

For a random sample $U \sim \mathcal{P}$, denote $\mathbb{E}[\|U\|^2]$ as $H(\mathcal{P})$.

Recall that each iteration involves two sources of randomness stemming from g_z and $\{U_i^{(j)}\}_{j=1}^k$, respectively. Let us define

$$\delta_i = \frac{1}{2k}\sum_{j=1}^{k}\left(\nabla f\big(W_i + U_i^{(j)}\big) + \nabla f\big(W_i - U_i^{(j)}\big)\right) - \nabla F(W_i),$$
$$\xi_i = \frac{1}{2k}\sum_{j=1}^{k}\left(G_i^{(j)} - \nabla f\big(W_i + U_i^{(j)}\big) - \nabla f\big(W_i - U_i^{(j)}\big)\right),$$

for $i = 0, \ldots, T-1$. One can see that both δ_i and ξ_i have mean zero. The former is by the symmetry of $\mathcal{P}$. The latter is because g_z is unbiased under Assumption 4.1. The next result gives their variance.

Lemma 4.4. *In the setting of Theorem 4.3, for any $i = 1, \ldots, T$, we have*

$$\mathbb{E}\left[\|\xi_i\|^2\right] \leq \frac{\sigma^2}{k} \quad \textit{and} \quad \mathbb{E}\left[\|\delta_i\|^2\right] \leq \frac{C^2 H(\mathcal{P})}{k}. \tag{7}$$

The last step is using smoothness to show that $\|\nabla F(W_t)\|$ keeps reducing. For details, see Appendix B.1. As a remark, existing sharpness-reducing methods such as SAM (Foret et al., 2021) seem to suffer from issues of oscillation (Bartlett et al., 2023) around the local basin, leaving a convergence analysis challenging to achieve. By contrast, our approach can be analyzed with standard techniques from stochastic optimization (Ghadimi & Lan, 2013). This connection is not known before to our knowledge, and we believe is the key strength of our approach compared with existing sharpness-reducing methods.

Next, we construct an example to match the rate of the above analysis, essentially showing that the gradient norm bounds are tight (under the current assumptions). We use an example from the work of Drori & Shamir (2020). The difference here, in particular, is that we have to deal with the perturbations that have been added to the objective. For $t = 0, 1, \ldots, d-1$, let $e_t \in \mathbb{R}^d$ be the basis vector in dimension d, whose t-th coordinate is 1, while the remaining coordinates are all zero. Let $f : \mathbb{R}^d \to \mathbb{R}$ be defined as

$$f(W) = \frac{1}{2G}\langle W, e_0\rangle^2 + \sum_{i=0}^{T-1} h_i(\langle W, e_{i+1}\rangle), \tag{8}$$

where h_i is a piece-wise quadratic function parameterized by α_i, defined as follow:

$$h_i(x) = \begin{cases} \frac{C\alpha_i^2}{4} & |x| \leq \alpha_i, \\ -\frac{C\left(|x|-\alpha_i\right)^2}{2} + \frac{C\alpha_i^2}{4} & \alpha_i \leq |x| \leq \frac{3}{2}\alpha_i, \\ \frac{C\left(|x|-2\alpha_i\right)^2}{2} & \frac{3}{2}\alpha_i \leq |x| \leq 2\alpha_i, \\ 0 & 2\alpha_i \leq |x|. \end{cases}$$

One can verify that for each piece above, ∇h_i is C-Lipschitz. As a result, provided that $G \leq C^{-1}$, ∇f is C-Lipschitz, based on the definition of f in equation (8).

The stochastic function F requires setting the perturbation distribution $\mathcal{P}$. We set $\mathcal{P}$ by truncating an isotropic Gaussian $N(0, \sigma^2 \operatorname{Id}_d)$ so that the i-th coordinate is at most $2^{-1}\alpha_{i-1}$, for $i = 1, \ldots, T$. Additionally, we set the initialization W_0 to satisfy $\langle W_0, e_i\rangle = 0$ for any $i \geq 1$ while $\langle W_0, e_0\rangle \neq 0$. Finally, we choose the gradient oracle to satisfy that the i-th step's gradient noise $\xi_i = \langle \xi_i, e_{i+1}\rangle e_{i+1}$, which means that ξ_i is along the direction of the basis vector e_{i+1}. In particular, this implies only coordinate $i+1$ is updated in step i, as long as $\langle \xi_i, e_{i+1}\rangle \leq 2^{-1}\alpha_i$.

Theorem 4.5. *Let the learning rates $\eta_0, \ldots, \eta_{T-1}$ be at most C^{-1}. Let $D > 0$ be a fixed value. When they either satisfy $\sum_{i=0}^{T-1} \eta_i \lesssim \sqrt{kT}$, or $\eta_i = \eta < C^{-1}$ for any epoch i, then for the above construction, the following must hold*

$$\min_{1 \le t \le T} \mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \ge D\sqrt{\frac{C\sigma^2}{32k \cdot T}}. \tag{9}$$

We remark that the above construction requires $T \le d$. Notice that this is purely for technical reasons due to the construction. It is an interesting question whether this condition can be removed or not. We briefly illustrate the key ideas of the result. At step i, the gradient noise ξ_i plus the perturbation noise is less than $2^{-1}\alpha_i + 2^{-1}\alpha_i = \alpha_i$ at coordinate $i+1$ (by triangle inequality). Thus, $h_i'(\langle W_t, e_{i+1}\rangle) = 0$, which holds for all prior update steps. This implies

$$\nabla f(W_i) = G^{-1}\langle W_i, e_0\rangle.$$

Recall from Assumption 4.2 that $F(W_0) \le D^2$. This condition imposes how large the α_i's can be. In particular, in the proof we will set $\alpha_i = 2\eta_i\sigma/\sqrt{k}$. Then, based on the definition of $f(W_0)$,

$$h_i(\langle W_0, e_{i+1}\rangle) = \frac{C\alpha_i^2}{4}, \text{ since } \langle W_0 + U, e_{i+1}\rangle \le \alpha_i.$$

In Lemma B.2, we then argue that the learning rates in this case must satisfy $\sum_{i=0}^{T-1} \eta_i \le O(\sqrt{T})$.

When the learning rate is fixed and at least $\Omega(T^{-1/2})$, we construct a piece-wise quadratic function (similar to equation (8)), now with a fixed α. This is described in Lemma B.3. In this case, the gradient noise grows by $1 - C^{-1}\eta$ up to T steps. We then carefully set α to lower bound the norm of the gradient. Combining these two cases, we conclude the proof of Theorem 4.5. For details, see Appendix B.2. As is typical in lower-bound constructions, our result holds for a specific instance that covers a specific range of learning rates. It may be an interesting question to examine a broader range of instances for future work.

The proof can also be extended to adaptive learning rate schedules. Notice that the above construction holds for arbitrary learning rates defined as a function of previous iterates. Then, we set the width of each function h_t, α_t, proportional to $\eta_t > 0$, for any η_t that may depend on previous iterates, as long as they satisfy the constraint that $\sum_{i=0}^{T-1} \eta_i \le O(\sqrt{T})$.

We can show a similar lower bound for the momentum update rule. Recall this is defined as

$$M_{i+1} = \mu M_i - \eta_i G_i, \text{ and } W_{i+1} = W_i + M_{i+1}, \tag{10}$$

for $i = 0, 1, \ldots, T-1$, where G_i is the specific gradient at step i. To handle this case, we will need a more fine-grained control on the gradient, so we consider a quadratic function as $f(W) = \frac{C}{2}\|W\|^2$. We leave the result and its proof to Appendix B.3.

5 Conclusion

This paper examines the injection of noise into the weights of a neural network. We begin by observing that the natural approach of injecting noise into the weight before running SGD does not work well in practice. Through extensive experiments for fine-tuning pretrained models, we show that a two-point noise injection method can indeed regularize the Hessian effectively, improving upon SGD, perturbed SGD, and SAM. Moreover, we show a generalization bound for model fine-tuning using PAC-Bayes analysis. Compared with four sharpness-reducing methods, our proposed algorithm yields statistically significant improvements across a wide range of data sets. Lastly, we provide a convergence analysis of the proposed algorithm.

We discuss several avenues for future work. Can the newly developed techniques be used for studying transformer networks? Can we better understand the dynamics of the Hessian during training? More broadly, there are many promising directions to explore, by approaching generalization through measuring geometric properties of large models.

References

Pierre Alquier. User-friendly introduction to pac-bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021. 6

Guozhong An. The effects of adding noise during backpropagation training on a generalization performance. *Neural computation*, 8(3):643–674, 1996. 2, 3

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *ICML*, 2022. 1

Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. *Advances in Neural Information Processing Systems*, 36, 2024. 2

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023. 3

Francis Bach. Learning theory from first principles. *Online version*, 2021. 23

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. 1, 3, 12

Devansh Bisla, Jing Wang, and Anna Choromanska. Low-pass filtering sgd for recovering flat optima in the deep learning optimization landscape. In *International Conference on Artificial Intelligence and Statistics*, pp. 8299–8339. PMLR, 2022. 3

Alexander Camuto, Matthew Willetts, Umut Simsekli, Stephen J Roberts, and Chris C Holmes. Explicit regularisation in gaussian noise injections. *Advances in Neural Information Processing Systems*, 33:16603–16614, 2020. 3

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120, 2020. 2, 3

Olivier Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007. 2, 6

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *ICLR*, 2021. 3

Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Symposium on Theory of Computing*, 2021. 3

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017. 1

Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *ICML*, 2020. 2, 12

John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 2015. 4, 11

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *UAI*, 2017. 2

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in pac-bayes bounds. In *International Conference on Artificial Intelligence and Statistics*, pp. 604–612. PMLR, 2021. 2

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *ICLR*, 2021. 1, 9, 12

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. 2, 11, 12, 23

Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011. 2, 3

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 4

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13, 1993. 2, 3

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

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *UAI*, 2018. 1

Haotian Ju, Dongyue Li, and Hongyang R Zhang. Robust fine-tuning of deep neural networks with hessian-based generalization guarantees. *ICML*, 2022. 2

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. *ICLR*, 2017. 1

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *ICML*, 2021. 9

Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020. 2, 11

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pp. 2–47. PMLR, 2018. 6

Yong Liu, Siqi Mai, Minhao Cheng, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Random sharpness-aware minimization. *Advances in Neural Information Processing Systems*, 2022. 9

Philip M Long and Peter L Bartlett. Sharpness-aware minimization and the edge of stability. *arXiv preprint arXiv:2309.12488*, 2023. 3

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 4

David McAllester. A pac-bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307.2118*, 2013. 2, 6, 17

David A McAllester. Some pac-bayesian theorems. *Machine Learning*, 1999. 3

Thomas Möllenhoff and Mohammad Emtiyaz Khan. Sam as an optimal relaxation of bayes. In *International Conference on Learning Representations*, 2023. 9

Vaishnavh Nagarajan and J Zico Kolter. Deterministic pac-bayesian generalization bounds for deep networks via generalizing noise-resilience. *ICLR*, 2020. 1

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017. 3

Antonio Orvieto, Anant Raj, Hans Kersting, and Francis Bach. Explicit regularization in overparametrized models via noise injection. *AISTATS*, 2023. 2, 3, 4

Samiksha Pachade, Prasanna Porwal, Dhanshree Thulkar, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Luca Giancardo, Gwenolé Quellec, and Fabrice Mériaudeau. Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. *Data*, 6(2):14, 2021. 4

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010. 7, 21

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. 5

John Shawe-Taylor and Robert C Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pp. 2–9, 1997. 3

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International conference on computational learning theory*, pp. 545–560. Springer, 2005. 5

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis. In *International Conference on Machine Learning*, pp. 9636–9647. PMLR, 2020. 2

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019. 20, 21

Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How does sharpness-aware minimization minimize sharpness? *ICLR*, 2023. 1, 2

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022a. 1, 3, 4

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7959–7971, 2022b. 3

Tong Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023. 2, 11

A Omitted Proofs from Section 2

Notations: We state a few standard notations first. Given two matrices X, Y having the same dimension, let $\langle X, Y\rangle = \mathrm{Tr}[X^\top Y]$ denote the matrix inner product of X and Y. Let $\|X\|_2$ denote the spectral norm (largest singular value) of X, and let $\|X\|_F$ denote the Frobenius norm of X. We use the big-O notation $f(x) = O(g(x))$ to indicate that there exists a fixed constant C independent of x such that $f(x) \leq C \cdot g(x)$ for large enough values of x.

A.1 Proof of Hessian-based PAC-Bayes Bound

We will use the following PAC-Bayes bound (for reference, see, e.g., Theorem 2, McAllester (2013)).

Theorem A.1. *Suppose the loss function $\ell(f_W(x), y)$ lies in a bounded range $[0, C]$ given any $x \in \mathcal{X}$ with label y. For any $\beta \in (0, 1)$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following holds:*

$$L_{\mathcal{Q}}(W) \leq \frac{1}{\beta}\hat{L}_{\mathcal{Q}}(W) + \frac{C\big(KL(\mathcal{Q}||\mathcal{P}) + \log\frac{1}{\delta}\big)}{2\beta(1-\beta)n}. \tag{11}$$

This result provides flexibility in setting β. Our results will set β to balance the perturbation error of $\mathcal{Q}$ and the KL divergence between $\mathcal{P}$ and $\mathcal{Q}$. We will need the KL divergence between the prior $\mathcal{P}$ and the posterior $\mathcal{Q}$ in the PAC-Bayesian analysis. This is stated in the following result.

Proposition A.2. *Suppose $\mathcal{P} = N(X, \Sigma)$ and $\mathcal{Q} = N(Y, \Sigma)$ are both Gaussian distributions with mean vectors given by $X \in \mathbb{R}^p, Y \in \mathbb{R}^p$, and population covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. The KL divergence between $\mathcal{P}$ and $\mathcal{Q}$ is equal to*

$$KL(\mathcal{Q}||\mathcal{P}) = \frac{1}{2}(X - Y)^\top \Sigma^{-1}(X - Y).$$

Specifically, if $\Sigma = \sigma^2 \,\mathrm{Id}_p$, then the above simplifies to

$$KL(\mathcal{Q}||\mathcal{P}) = \frac{\|X - Y\|_2^2}{2\sigma^2}.$$

We will use Taylor's expansion on the perturbed loss. This is stated precisely as follows.

Claim A.3. *Let f_W be twice-differentiable, parameterized by weight vector $W \in \mathbb{R}^p$. Let $U \in \mathbb{R}^p$ be another vector with dimension p. For any W and U, the following identity holds*

$$\ell(f_{W+U}(x), y) = \ell(f_W(x), y) + U^\top \nabla \ell(f_W(x), y) + U^\top[\nabla^2 \ell(f_W(x), y)]U + R_2(\ell(f_W(x), y)),$$

where $R_2(\ell(f_W(x), y)))$ is a second-order error term in Taylor's expansion.

Proof. The proof follows by the fact that $\ell \circ f_W$ is twice-differentiable. From the mean value theorem, let $\eta \in \mathbb{R}^p$ be a vector that has the same dimension as W and U. There must exist an η between W and $U + W$ such that the following equality holds:

$$R_2(\ell(f_W(x), y)) = U^\top \Big(\nabla^2[\ell(f_\eta(x), y)] - \nabla^2[\ell(f_W(x), y)]\Big)U.$$

This completes the proof of the claim. □

Based on the above, we provide Taylor's expansion of the gap between $\ell_{\mathcal{Q}}$ and ℓ.

Lemma A.4. *In the setting of Theorem 2.2, suppose each parameter is perturbed by an independent noise drawn from $N(0, \sigma^2)$. Let $\ell_{\mathcal{Q}}(f_W(x), y)$ be the perturbed loss with noise perturbation injection vector on W. There exist some fixed value C_1 that do not grow with n and $1/\delta$ such that*

$$\left|\ell_{\mathcal{Q}}(f_W(x), y) - \ell(f_W(x), y) - \frac{1}{2}\sigma^2 \,\mathrm{Tr}\left[\nabla^2[\ell(f_W(x), y)]\right]\right| \leq C_1\sigma^3.$$

Proof. We take the expectation over U for both sides of the equation in Claim A.3. The result becomes

$$\mathop{\mathbb{E}}_{U}[\ell(f_{W+U}(x),y)] = \mathop{\mathbb{E}}_{U}\left[\ell(f_W(x),y) + U^\top \nabla \ell(f_W(x),y) + U^\top \nabla^2[\ell(f_W(x),y)]U + R_2(\ell(f_W(x),y))\right].$$

Then, we use the perturbation distribution $\mathcal{Q}$ on $\mathbb{E}_U[\ell(f_{W+U}(x),y)]$, and get

$$\ell_{\mathcal{Q}}(f_W(x),y) = \mathop{\mathbb{E}}_{U}[\ell(f_W(x),y)] + \mathop{\mathbb{E}}_{U}\left[U^\top \nabla \ell(f_W(x),y)\right] + \mathop{\mathbb{E}}_{U}\left[U^\top \nabla^2[\ell(f_W(x),y)]U\right] + \mathop{\mathbb{E}}_{U}[R_2(\ell(f_W(x),y))].$$

Since $\mathbb{E}[U]=0$, the first-order term will be zero in expectation. The second-order term becomes equal to

$$\mathop{\mathbb{E}}_{U}\left[U^\top[\nabla^2 \ell(f_W(x),y)]U\right] = \sigma^2 \operatorname{Tr}\left[\nabla^2[\ell(f_W(x),y)]\right]. \tag{12}$$

The expectation of the error term $R_2(\ell(f_W(x),y))$ be

$$\begin{aligned}
\mathop{\mathbb{E}}_{U}[R_2(\ell(f_W(x),y))] &= \mathop{\mathbb{E}}_{U}\left[U^\top\left(\nabla^2[\ell(f_\eta(x),y)] - \nabla^2[\ell(f_W(x),y)]\right)U\right] \\
&\le \mathop{\mathbb{E}}_{U}\left[\|U\|_2^2 \cdot \left\|\nabla^2[\ell(f_\eta(x),y)] - \nabla^2[\ell(f_W(x),y)]\right\|_F\right] \\
&\lesssim \mathop{\mathbb{E}}_{U}\left[\|U\|_2^2 \cdot C_1\|U\|_2\right] \lesssim C_1\sigma^3.
\end{aligned}$$

Thus, the proof is complete. □

The last piece we will need is the uniform convergence of the Hessian operator. The result uses the fact that the Hessian matrix is Lipschitz continuous.

Lemma A.5. *In the setting of Theorem 2.2, there exist some fixed values C_2, C_3 that do not grow with n and $1/\delta$, such that with probability at least $1-\delta$ for any $\delta > 0$, over the randomness of the n training examples, we have*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2[\ell(f_W(x_i),y_i)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\nabla^2[\ell(f_W(x),y)]\right]\right\|_F \le \frac{C_2\sqrt{\log(C_3 n/\delta)}}{\sqrt{n}}. \tag{13}$$

The proof will be deferred to Section A.1.2. With these results ready, we will now state the proof of the Hessian-based generalization bound.

A.1.1 Proof of Theorem 2.2

Proof of Theorem 2.2. First, we separate the gap of $L(W)$ and $\frac{1}{\beta}\hat{L}(W)$ into three parts:

$$L(W) - \frac{1}{\beta}\hat{L}(W) = L(W) - L_{\mathcal{Q}}(W) + L_{\mathcal{Q}}(W) - \frac{1}{\beta}\hat{L}_{\mathcal{Q}}(W) + \frac{1}{\beta}\hat{L}_{\mathcal{Q}}(W) - \frac{1}{\beta}\hat{L}(W).$$

By Lemma A.4, we can bound the difference between $L(W)$ and $L_{\mathcal{Q}}(W)$ by the Hessian trace plus an error:

$$\begin{aligned}
L(W) - \frac{1}{\beta}\hat{L}(W) \le &- \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\frac{\sigma^2}{2}\operatorname{Tr}\left[\nabla^2[\ell(f_W(x),y)]\right]\right] + C_1\sigma^3 + \left(L_{\mathcal{Q}}(W) - \frac{1}{\beta}\hat{L}_{\mathcal{Q}}(W)\right) \\
&+ \frac{1}{\beta}\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{\sigma^2}{2}\operatorname{Tr}\left[\nabla^2[\ell(f_W(x_i),y_i)]\right] + C_1\sigma^3\Big).
\end{aligned}$$

After re-arranging the terms, we can get the following:

$$\begin{aligned}
L(W) - \frac{1}{\beta}\hat{L}(W) \le &\underbrace{- \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\frac{\sigma^2}{2}\operatorname{Tr}\left[\nabla^2[\ell(f_W(x),y)]\right]\right] + \frac{1}{n\beta}\sum_{i=1}^{n}\frac{\sigma^2}{2}\operatorname{Tr}\left[\nabla^2[\ell(f_W(x_i),y_i)]\right]}_{E_1} \\
&+ \frac{1+\beta}{\beta}C_1\sigma^3 + \underbrace{L_{\mathcal{Q}}(W) - \frac{1}{\beta}\hat{L}_{\mathcal{Q}}(W)}_{E_2}.
\end{aligned} \tag{14}$$

We will examine E_1 by separating it into two parts:

$$E_1 = \frac{1}{\beta}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\sigma^2}{2}\operatorname{Tr}\left[\nabla^2[\ell(f_{\hat{W}}(x_i), y_i)]\right] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\frac{\sigma^2}{2}\operatorname{Tr}\left[\nabla^2[\ell(f_W(x), y)]\right]\right]\right) \tag{15}$$

$$+ \frac{1-\beta}{\beta}\frac{\sigma^2}{2}\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\operatorname{Tr}\left[\nabla^2\ell(f_W(x), y)\right]\right]. \tag{16}$$

We can use the uniform convergence result of Lemma A.5 to bound equation (15), leading to:

$$\begin{aligned}
&\frac{\sigma^2}{2\beta}\left(\frac{1}{n}\sum_{i=1}^{n}\operatorname{Tr}\left[\nabla^2\ell(f_W(x_i), y_i)\right] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\operatorname{Tr}\left[\nabla^2\ell(f_W(x), y))\right]\right]\right)\\
\leq &\frac{\sigma^2}{2\beta}\cdot\sqrt{p}\cdot\left\|\frac{1}{n}\sum_{i=1}^{n}\operatorname{Tr}\left[\nabla^2[\ell(f_W(x_i), y_i)]\right] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\operatorname{Tr}\left[\nabla^2[\ell(f_W(x), y)]\right]\right]\right\|_F \qquad \text{(by Cauchy-Schwarz)}\\
\leq &\frac{\sigma^2\sqrt{p}\cdot C_2\sqrt{\log(C_3 n/\delta)}}{2\beta\sqrt{n}}.
\end{aligned} \tag{17}$$

As for equation (16), we recall that

$$\alpha := \max_{(x,y)\sim\mathcal{D}}\operatorname{Tr}\left[\nabla^2\ell(f_W(x), y)\right].$$

Combined with equation (17), we have shown that

$$E_1 \leq \frac{\sigma^2\sqrt{p}\cdot C_2\sqrt{\log(C_3 n/\delta)}}{2\beta\sqrt{n}} + \frac{1-\beta}{\beta}\frac{\sigma^2}{2}\cdot\alpha. \tag{18}$$

As for E_2, we will use the PAC-Bayes bound of Theorem A.1. In particular, we set the prior distribution $\mathcal{P}$ as the distribution of U and we set the posterior distribution $\mathcal{Q}$ as the distribution of $W+U$. Thus,

$$E_2 \leq \frac{C\big(KL(\mathcal{Q}||\mathcal{P}) + \log\frac{1}{\delta}\big)}{2\beta(1-\beta)n} \leq \frac{C\Big(\frac{\|W\|_2^2}{2\sigma^2} + \log\frac{1}{\delta}\Big)}{2\beta(1-\beta)n} \leq \frac{C(\frac{r^2}{2\sigma^2} + \log\delta^{-1})}{2\beta(1-\beta)n}. \tag{19}$$

The last step is because $\|W\|_2 \leq r$ by assumption of the hypothesis space. Combining equations (14), (18), (19), we claim that with probability at least $1-2\delta$, the following must be true:

$$L(W) - \frac{1}{\beta}\hat{L}(W) \leq \frac{\sigma^2\sqrt{p}\cdot C_2\sqrt{\log(C_3 n/\delta)}}{2\beta\sqrt{n}} + \frac{1-\beta}{\beta}\frac{\sigma^2}{2}\alpha + \frac{1+\beta}{\beta}C_1\sigma^3 + \frac{C(\frac{r^2}{2\sigma^2} + \log\frac{1}{\delta})}{2\beta(1-\beta)n}. \tag{20}$$

Thus, we will now choose σ and $\beta \in (0,1)$ to minimize the term above. In particular, we will set σ such that:

$$\sigma^2 = \frac{r}{1-\beta}\sqrt{\frac{C}{\alpha n}}. \tag{21}$$

By plugging in this setting to equation (20) and re-arranging terms, the gap between $L(W)$ and $\hat{L}(W)/\beta$ becomes:

$$L(W) - \frac{1}{\beta}\hat{L}(W) \leq \frac{1}{\beta}\sqrt{\frac{C\alpha r^2}{n}} + \frac{C_2\sqrt{2p\log(C_3 n/\delta)}}{2\beta\sqrt{n}}\sigma^2 + \frac{1+\beta}{\beta}C_1\sigma^3 + \frac{C}{2\beta(1-\beta)n}\log\frac{1}{\delta}.$$

Let β be a fixed value close to 1 and independent of N and δ^{-1}, and let $\epsilon = (1-\beta)/\beta$. We get

$$L(W) \leq (1+\epsilon)\hat{L}(W) + (1+\epsilon)\sqrt{\frac{C\alpha r^2}{n}} + \xi, \text{ where}$$

$$\xi = \frac{C_2\sqrt{2p\log(C_3 n/\delta)}}{2\beta\sqrt{n}}\sigma^2 + \Big(1+\frac{1}{\beta}\Big)C_1\sigma^3 + \frac{C}{2\beta(1-\beta)n}\log\frac{1}{\delta}.$$

Notice that ξ is of order $O(n^{-\frac{3}{4}} + n^{-\frac{3}{4}} + \log(\delta^{-1})n^{-1}) \leq O(\log(\delta^{-1})n^{-\frac{3}{4}})$. Therefore, we have finished the proof of equation (2). □

A.1.2 Proof of Lemma A.5

In this section, we provide the proof of Lemma A.5, which shows the uniform convergence of the loss Hessian.

Proof of Lemma A.5. Let C, $\epsilon > 0$, and let $S = \{W \in \mathbb{R}^p : \|W\|_2 \leq C\}$. There exists an ϵ-cover of S with respect to the ℓ_2-norm at most $\max\left(\left(\frac{3C}{\epsilon}\right)^p, 1\right)$ elements; see, e.g., Example 5.8 (Wainwright, 2019). Let $T \subseteq S$ denote the set of this cover. Recall that the Hessian $\nabla^2[\ell(f_W(x), y)]$ is C_1-Lipschitz for all $(W + U) \in S, W \in S$. Then we have

$$\left\|\nabla^2[\ell(f_{W+U}(x), y)] - \nabla^2[\ell(f_W(x), y)]\right\|_F \leq C_1 \|U\|_2 .$$

For parameters $\delta, \epsilon > 0$, let $\mathcal{N}$ be the ϵ-cover of S with respect to the ℓ_2-norm. Define the event

$$E = \left\{\forall W \in T, \left\|\frac{1}{n}\sum_{i=1}^{n} \nabla^2[\ell(f_W(x_i), y_i)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\nabla^2[\ell(f_W(x), y)]\right]\right\|_F \leq \delta\right\}.$$

By the matrix Bernstein inequality, we have

$$\Pr[E] \geq 1 - 4 \cdot |\mathcal{N}| \cdot p \cdot \exp\left(-\frac{n\delta^2}{2\alpha^2}\right).$$

Next, for any $W \in S$, we can pick some $W + U \in T$ such that $\|U\|_2 \leq \epsilon$. We have

$$\left\|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\nabla^2[\ell(f_{W+U}(x), y)]\right] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\nabla^2[\ell(f_W(x), y)]\right]\right\|_F \leq C_1 \|U\|_2 \leq C_1\epsilon$$

$$\left\|\frac{1}{n}\sum_{j=1}^{n} \nabla^2[\ell(f_{W+U}(x_j), y_j)] - \frac{1}{n}\sum_{j=1}^{n} \nabla^2[\ell(f_W(x_j), y_j)]\right\|_F \leq C_1 \|U\|_2 \leq C_1\epsilon.$$

Therefore, for any $W \in S$, we obtain:

$$\left\|\frac{1}{n}\sum_{j=1}^{n} \nabla^2[\ell(f_W(x_j), y_j)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\nabla^2[\ell(f_W(x), y)]\right]\right\|_F \leq 2C_1\epsilon + \delta.$$

We will also set the value of δ and ϵ. First, set $\epsilon = \delta/(2C_1)$ so that conditional on E,

$$\left\|\frac{1}{n}\sum_{j=1}^{n} \nabla^2[\ell(f_W(x_j), y_j)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\nabla^2[\ell(f_W(x), y)]\right]\right\|_F \leq 2\delta.$$

The event E happens with a probability of at least:

$$1 - 4|T|p \cdot \exp\left(-\frac{n\delta^2}{2\alpha^2}\right) = 1 - 4p \cdot \exp\left(\log|T| - \frac{n\delta^2}{2\alpha^2}\right).$$

We have $\log|T| \leq p\log(3B/\epsilon) = p\log(6CC_1/\delta)$. If we set

$$\delta = \sqrt{\frac{4p\alpha^2\log(3\tau CC_1 n/\alpha)}{n}}$$

so that $\log(3\tau CC_1 n/\alpha) \geq 1$ (because $n \geq \frac{e\alpha}{3C_1}$ and $\tau \geq 1$), then we get

$$\begin{aligned}
p\log(6CC_1/\delta) - n\delta^2/(2\alpha^2) =& p\log\left(\frac{6CC_1\sqrt{n}}{\sqrt{4p\alpha^2\log(3\tau CC_1 n/\alpha)}}\right) - 2p\log(3\tau CC_1 n/\alpha) \\
=& p\log\left(\frac{3CC_1\sqrt{n}}{\alpha\sqrt{p\log(3\tau CC_1 n/\alpha)}}\right) - 2p\log(3\tau CC_1 n/\alpha) \\
\leq& p\log(3\tau CC_1 n/\alpha) - 2p\log(3\tau CC_1 n/\alpha) \qquad (\tau \geq 1, \log(3\tau CC_1 n/\alpha) \geq 1) \\
=& -p\log(3\tau CC_1 n/\alpha) \leq -p\log(e\tau). \qquad (3CC_1 n/\alpha \geq e)
\end{aligned}$$

Therefore, with a probability greater than

$$1 - 4|\mathcal{N}|p \cdot \exp(-n\delta^2/(2\alpha^2)) \geq 1 - 4p(e\tau)^{-p},$$

the following estimate holds:

$$\left\| \frac{1}{n} \sum_{j=1}^{n} \nabla^2[\ell(f_W(x_j), y_j)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\nabla^2[\ell(f_W(x), y)]\right] \right\|_F \leq \sqrt{\frac{16p\alpha^2 \log(3\tau C C_1 n/\alpha)}{n}}.$$

Denote $\delta' = 4p(e\tau)^{-p}$, $C_2 = 4\alpha\sqrt{p}$, and $C_3 = 12pCC_1/(e\alpha)$. With probability greater than $1 - \delta'$, the final result is:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla^2[\ell(f_W(x_i), y_i)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\nabla^2[\ell(f_W(x), y)]\right] \right\|_F \leq C_2 \sqrt{\frac{\log(C_3 n/\delta')}{n}}.$$

This completes the proof of Lemma A.5. □

A.2 Proof of Proposition 2.3

Proof of Proposition 2.3. We can calculate the gradient as

$$\nabla \hat{L}(W) = \frac{1}{n} \sum_{i=1}^{n} (\langle A_i, WW^\top \rangle - y_i) A_i W. \tag{22}$$

For a particular entry $W_{j,k}$ of W, for any $1 \leq j, k \leq d$, the derivative of the above gradient with respect to $W_{j,k}$ is

$$\frac{1}{n} \sum_{i=1}^{n} \left([A_i W]_{j,k} A_i W + \left(\langle A_i, WW^\top \rangle - y_i \right) \frac{\partial (A_i W)}{\partial W_{j,k}} \right). \tag{23}$$

When $\hat{L}(W)$ is zero, the second term of equation (23) above must be zero, because $\langle A_i, WW^\top \rangle$ is equal to y_i, for any $i = 1, \ldots, n$.

Now, we use the assumption that A_i is a random Gaussian matrix, in which every entry is drawn from a normal distribution with mean zero and variance one. Notice that the expectation of $\|A_i W\|_F^2$ satisfies:

$$\mathbb{E}\left[\|A_i W\|_F^2\right] = \mathbb{E}\left[\mathrm{Tr}\left[W^\top A_i^\top A_i W\right]\right] = \mathrm{Tr}\left[W^\top (d \cdot \mathrm{Id}_{d\times d}) W^\top\right] = d \cdot \mathrm{Tr}\left[W^\top W\right] = d \|W\|_F^2.$$

Thus, by concentration inequality for χ^2 random variables (e.g., Wainwright (2019, equation (2.19))), the following holds for any $0 < \epsilon < 1$,

$$\Pr\left[\left| \frac{1}{n} \sum_{i=1}^{n} \|A_i W\|_F^2 - d\|W\|_F^2 \right| \geq \epsilon d \|W\|_F^2 \right] \leq 2\exp\left(-\frac{n\epsilon^2}{8}\right). \tag{24}$$

This implies that ϵ must be smaller than $O(n^{-1/2})$ with high probability. As a result, the average of $\|A_i W\|_F^2$ must be $d\|W\|_F^2$ plus some deviation error that scales with $n^{-1/2}$ times the expectation.

By Theorem 3.2, Recht et al. (2010), the minimum Frobenius norm ($\|W\|_F^2$) solution that satisfies $\hat{L}(W) = 0$ (for Gaussian random matrices) is precisely $U^\star$. Thus, we conclude that equation (5) holds. □

B Omitted Proofs from Section 4

B.1 Proof of Theorem 4.3

First, let us show that ∇F is C-Lipschitz. To see this, we apply the Lipschitz condition of the gradient inside the expectation of $F(W)$. For any $W_1, W_2 \in \mathbb{R}^d$, by definition,

$$\begin{aligned}\|\nabla F(W_1) - \nabla F(W_2)\| &= \left\|\nabla \mathop{\mathbb{E}}_{U\sim\mathcal{P}}[f(W_1+U)] - \nabla \mathop{\mathbb{E}}_{U\sim\mathcal{P}}[f(W_2+U)]\right\| \\ &= \left\|\mathop{\mathbb{E}}_{U\sim\mathcal{P}}[\nabla f(W_1+U) - \nabla f(W_2+U)]\right\| \\ &\leq \mathop{\mathbb{E}}_{U\sim\mathcal{P}}[\|\nabla f(W_1+U) - \nabla f(W_2+U)\|] \leq C\,\|W_1 - W_2\|\,.\end{aligned}$$

Next, we provide the proof for bounding the variance of δ_i and ξ_i for $i = 0, 1, \ldots, T-1$.

Proof. First, we can see that

$$\begin{aligned}\mathop{\mathbb{E}}_{U_i^1,\ldots,U_i^k}\left[\|\delta_i\|^2\right] &= \mathop{\mathbb{E}}_{U_i^1,\ldots,U_i^k}\left[\left\|\frac{1}{2k}\sum_{j=1}^{k}\Big(\nabla f(W_i+U_i^j) + \nabla f(W_i-U_i^j) - 2\nabla F(W_i)\Big)\right\|^2\right] \\ &= \frac{1}{k^2}\sum_{j=1}^{k}\mathop{\mathbb{E}}_{U_i^j}\left[\left\|\frac{1}{2}\Big(\nabla f(W_i+U_i^j) + \nabla f(W_i-U_i^j) - 2\nabla F(W_i)\Big)\right\|^2\right] \qquad (25) \\ &= \frac{1}{k}\mathop{\mathbb{E}}_{U_i^1}\left[\left\|\frac{1}{2}\Big(\nabla f(W_i+U_i^1) + \nabla f(W_i-U_i^1)\Big) - \nabla F(W_i)\right\|^2\right] \qquad (26)\end{aligned}$$

where in the second line we use that $U_i^{j_1}$ and $U_i^{j_2}$ are independent when $j_1 \neq j_2$, in the last line we use fact that $U_i^1, \ldots, U_i^k$ are identically distributed. In the second step, we use the fact that for two independent random variables U, V, and any continuous functions $h(U), g(V)$, $h(U)$ and $g(V)$ are still independent (recall that f is continuous since it is twice-differentiable). We include a short proof of this fact for completeness. If U and V are independent, we have $\Pr[U \in A, V \in B] = \Pr[U \in A] \cdot \Pr[V \in B]$, for any $A, B \in \text{Borel}(\mathbb{R})$. Thus, if h and g are continuous functions, we obtain

$$\begin{aligned}\Pr[h(U) \in A, g(V) \in B] &= \Pr[U \in h^{-1}(A), V \in g^{-1}(B)] \\ &= \Pr[U \in h^{-1}(A)] \cdot \Pr[V \in g^{-1}(B)] = \Pr[h(U) \in A] \cdot \Pr[g(V) \in B].\end{aligned}$$

Thus, we have shown that

$$\mathbb{E}\left[\|\delta_i\|^2\right] = \frac{1}{k}\mathop{\mathbb{E}}_{U\sim\mathcal{P}}\left[\left\|\frac{1}{2}\Big(\nabla f(W_i+U) + f(W_i-U)\Big) - \nabla F(W_i)\right\|^2\right]. \qquad (27)$$

Next, we deal with the variance of the two-point stochastic gradient. We will show that

$$\mathop{\mathbb{E}}_{U}\left[\left\|\frac{1}{2}\Big(\nabla f(W+U) + \nabla f(W-U)\Big) - \nabla F(W)\right\|^2\right] \leq C^2 H(\mathcal{P}). \qquad (28)$$

We mainly use the Lipschitz continuity of the gradient of F. The left-hand side of equation (28) is equal to

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{U}\left[\left\|\frac{1}{2}\Big(\nabla f(W+U)-\nabla F(W)\Big)+\frac{1}{2}\Big(\nabla f(W-U)-\nabla F(W)\Big)\right\|^2\right]\\
\leq&\mathop{\mathbb{E}}_{U}\left[\frac{1}{2}\left\|\nabla f(W+U)-\nabla F(W)\right\|^2+\frac{1}{2}\left\|\nabla f(W-U)-\nabla F(W)\right\|^2\right] && \text{(by Cauchy-Schwartz)}\\
=&\frac{1}{2}\mathop{\mathbb{E}}_{U}\left[\left\|\nabla f(W+U)-\nabla F(W)\right\|^2\right] && \text{(by symmetry of } \mathcal{P} \text{ since it has mean zero)}\\
=&\frac{1}{2}\mathop{\mathbb{E}}_{U}\left[\left\|\mathop{\mathbb{E}}_{U'\sim\mathcal{P}}[\nabla f(W+U)-\nabla f(W+U')]\right\|^2\right]\\
\leq&\frac{1}{2}\mathop{\mathbb{E}}_{U}\left[\mathop{\mathbb{E}}_{U'\sim\mathcal{P}}\left[\left\|\nabla f(W+U)-\nabla f(W+U')\right\|^2\right]\right]\\
\leq&\frac{1}{2}\mathop{\mathbb{E}}_{U,U'}\left[C^2\left\|U-U'\right\|^2\right]=\frac{1}{2}C^2\mathop{\mathbb{E}}_{U,U'}\left[\left\|U\right\|^2+\left\|U'\right\|^2\right]=C^2H(\mathcal{P}) && \text{(by equation (30))}
\end{aligned}
$$

As for the variance of ξ_i, we note that $U_i^{(1)},\ldots,U_i^{(j)}$ are all independent from each other. Therefore,

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\left\{U_i^{(j)},z_i^{(j)}\right\}_{j=1}^k}\left[\left\|\xi_i\right\|^2\right]=&\frac{1}{4k}\mathop{\mathbb{E}}_{U,z}\left[\left\|g_z(W+U)-\nabla f(W+U)+g_z(W-U)-f(W-U)\right\|^2\right]\\
\leq&\frac{1}{2k}\mathop{\mathbb{E}}_{U,z}\left[\left\|g_z(W+U)-\nabla f(W+U)\right\|^2+\left\|g_z(W-U)-\nabla f(W-U)\right\|^2\right]\\
\leq&\frac{\sigma^2}{k}.
\end{aligned}
$$

The first step uses the fact that both $g_z(\cdot)$ and $f(\cdot)$ are continuous functions The second step above uses Cauchy-Schwartz inequality. The last step uses the variance bound of $g_z(\cdot)$, Thus, the proof is finished. □

Next, we show the convergence of the gradient, which is based on the classical work of Ghadimi & Lan (2013).

Lemma B.1. *In the setting of Theorem 4.3, for any $\eta_0,\cdots,\eta_{T-1}$ less than C^{-1} and a random variable according to a distribution $\Pr[t=j]=\frac{\eta_j}{\sum_{i=0}^{T-1}\eta_i}$, for any $j=0,\ldots,T-1$, the following holds:*

$$
\mathbb{E}\left[\left\|\nabla F(W_t)\right\|^2\right]\leq\frac{2C}{\sum_{i=0}^{T-1}\eta_i}D^2+\frac{C\sum_{i=0}^{T-1}\eta_i^2\big(\mathbb{E}\left[\left\|\delta_i\right\|^2\right]+\mathbb{E}\left[\left\|\xi_i\right\|^2\right]\big)}{\sum_{i=0}^{T-1}\eta_i}. \tag{29}
$$

Proof. The smoothness condition in Assumption 4.2 implies the following domination inequality:

$$
\left|F(W_2)-F(W_1)-\langle\nabla F(W_1),W_2-W_1\rangle\right|\leq\frac{C}{2}\left\|W_2-W_1\right\|^2. \tag{30}
$$

See, e.g., Bach (2021, Chapter 5). Here, we use the fact that $\nabla F(W)$ is L-Lipschitz continuous. Based on the above smoothness inequality, we have

$$
\begin{aligned}
&F(W_{i+1})\\
\leq&F(W_i)+\langle\nabla F(W_i),W_{i+1}-W_i\rangle+\frac{C}{2}\eta_i^2\left\|\frac{1}{2}\Big(\nabla f(W_i+U_i)+\nabla f(W_i-U_i)\Big)+\xi_i\right\|^2\\
=&F(W_i)-\eta_i\langle\nabla F(W_i),\delta_i+\xi_i+\nabla F(W_i)\rangle+\frac{C\eta_i^2}{2}\left\|\delta_i+\xi_i+\nabla F(W_i)\right\|^2\\
=&F(W_i)-\left(\eta_i-\frac{C\eta_i^2}{2}\right)\left\|\nabla F(W_i)\right\|^2-\left(\eta_i-C\eta_i^2\right)\langle\nabla F(W_i),\delta_i+\xi_i\rangle+\frac{C\eta_i^2}{2}\left\|\delta_i+\xi_i\right\|^2.
\end{aligned}
$$

Summing up the above inequalities for $i = 0, 1, \ldots, T-1$, we obtain

$$\sum_{i=0}^{T-1} F(W_{i+1}) \leq \sum_{i=0}^{T-1} F(W_i) - \sum_{i=0}^{T-1} \left(\eta_i - \frac{C\eta_i^2}{2}\right) \|\nabla F(W_i)\|^2 - \sum_{i=0}^{T-1} \left(\eta_i - C\eta_i^2\right) \langle \nabla F(W_i), \delta_i + \xi_i \rangle + \sum_{i=0}^{T-1} \frac{C\eta_i^2}{2} \|\delta_i + \xi_i\|^2,$$

which implies that

$$\sum_{i=0}^{T-1} \left(\eta_i - \frac{C\eta_i^2}{2}\right) \|\nabla F(W_i)\|^2 \tag{31}$$

$$\leq F(W_0) - F(W_T) - \sum_{i=0}^{T-1} \left(\eta_i - C\eta_i^2\right) \langle \nabla F(W_i), \delta_i + \xi_i \rangle + \frac{C}{2} \sum_{i=0}^{T-1} \eta_i^2 \|\delta_i + \xi_i\|^2$$

$$\leq D^2 - \sum_{i=0}^{T-1} \left(\eta_i - C\eta_i^2\right) \langle \nabla F(W_i), \delta_i + \xi_i \rangle + \frac{C}{2} \sum_{i=0}^{T-1} \eta_i^2 \|\delta_i + \xi_i\|^2. \tag{32}$$

where in the last step, we use the fact that

$$F(W_0) - F(W_T) \leq F(W_0) - \min_{W \in \mathbb{R}^d} F(W) \leq D^2.$$

For any $t = 0, 1, \ldots, T-1$, notice that as long as $0 < \eta_t \leq \frac{1}{C}$, then

$$\eta_t \leq 2\eta_t - C\eta_t^2.$$

Hence, we have

$$\frac{1}{2} \sum_{t=0}^{T-1} \eta_t \|\nabla F(W_t)\|^2 \leq \sum_{t=0}^{T-1} \left(\eta_t - \frac{C\eta_t^2}{2}\right) \|\nabla F(W_t)\|^2,$$

which implies that

$$\frac{1}{2} \sum_{i=0}^{T-1} \eta_i \|\nabla F(W_i)\|^2 \leq D^2 - \sum_{i=0}^{T-1} \left(\eta_i - C\eta_i^2\right) \langle \nabla F(W_i), \delta_i + \xi_i \rangle + \frac{C}{2} \sum_{i=0}^{T-1} \eta_i^2 \|\delta_i + \xi_i\|^2. \tag{33}$$

Additionally, since U_t is drawn from a distribution with mean zero. Hence, by symmetry, we get that

$$\mathop{\mathbb{E}}_{U_t}[\delta_t] = \frac{1}{2} \mathop{\mathbb{E}}_{U_t}[\nabla f(W_t - U_t) - \nabla f(W_t + U_t)] = 0. \tag{34}$$

Thus, if we take the expectation over $U_0, U_1, \ldots, U_{T-1}, \xi_0, \xi_1, \ldots, \xi_{T-1}$, then

$$\mathbb{E}[\langle \nabla F(W_i), \delta_i + \xi_i \rangle] = 0.$$

Recall that t is a random variable whose probability mass is specified in Lemma B.1. We can write equation (33) equivalently as (below, we take expectation over all the random variables along the update since W_t is a function of the previous gradient updates, for each $t = 0, 1, \ldots, T-1$, recalling that $\Pr[t = i] = \frac{\eta_i}{\sum_{j=0}^{T-1} \eta_j}$)

$$\begin{aligned}
\mathop{\mathbb{E}}_{t;\ U_0,\ldots,U_{T-1},\xi_0,\xi_1,\ldots,\xi_{T-1}} \left[\|\nabla F(W_t)\|^2\right] &= \frac{\sum_{i=0}^{T-1} \eta_i \mathbb{E}\left[\|\nabla F(W_i)\|^2\right]}{\sum_{i=0}^{T-1} \eta_i} \\
&\leq \frac{2D^2 + C\sum_{i=0}^{T-1} \eta_i^2 \mathbb{E}\left[\|\delta_i + \xi_i\|^2\right]}{\sum_{i=0}^{T-1} \eta_i} \\
&= \frac{2D^2 + C\sum_{i=0}^{T-1} \eta_i^2 \left(\mathbb{E}\left[\|\delta_i\|^2\right] + \mathbb{E}\left[\|\xi_i\|^2\right]\right)}{\sum_{i=0}^{T-1} \eta_i}.
\end{aligned}$$

where we use the fact that δ_i and ξ_i are independent for any i. Hence, we have finished the proof of equation (29). □

Based on the above result, we now finish the proof of the upper bound in Proposition 4.3.

Proof. Let the step sizes be equal to a fixed η for all epochs. Thus, Eq. (29) becomes

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \leq \frac{2}{T\eta}D^2 + \frac{C\eta}{T}\sum_{i=0}^{T-1}\left(\mathbb{E}\left[\|\delta_i\|^2\right] + \mathbb{E}\left[\|\xi_i\|^2\right]\right). \tag{35}$$

By Lemma 4.4,

$$\sum_{i=0}^{T-1}\left(\mathbb{E}\left[\|\delta_i\|^2\right] + \mathbb{E}\left[\|\xi_i\|^2\right]\right) \leq T \cdot \frac{\sigma^2 + C^2 H(\mathcal{P})}{k}. \tag{36}$$

For simplicity, let us denote $\Delta = \frac{\sigma^2 + C^2 H(\mathcal{P})}{k}$. The proof is divided into two cases.

Case 1: Δ is large. More precisely, suppose that $\Delta \geq 2CD^2/T$. Then, minimizing over η above leads us to the following upper bound on the right-hand side of equation (35):

$$\sqrt{\frac{2CD^2\Delta}{T}}, \tag{37}$$

which is obtained by setting

$$\eta = \sqrt{\frac{2D^2}{C\Delta T}}.$$

One can verify that this step size is less than $\frac{1}{C}$ since Δ is at least $2CD^2$. Thus, we conclude that equation (35) must be less than

$$\sqrt{\frac{2CD^2\Delta}{T}} = \sqrt{\frac{2CD^2(\sigma^2 + C^2 H(\mathcal{P})))}{kT}}. \tag{38}$$

Case 2: Δ is small. In this case, suppose $\Delta < 2CD^2/T$. Then, the right-hand side of equation (35) must be less than

$$\frac{2D^2}{T\eta} + \frac{2C^2D^2\eta}{T} \leq \frac{2CD^2}{T}. \tag{39}$$

Thus, combining equations (38) and (39), we have completed the proof of equation (6).

□

B.2 Proof of Theorem 4.5

Recall our construction from Section 4 as follows. Let e_t be the basis vector for the t-th dimension, for $t = 0, 1, \ldots, T-1$. Define $f(W)$ as

$$f(W) = \frac{1}{2G}\langle W, e_0\rangle^2 + \sum_{i=0}^{T-1} h_i(\langle W, e_{i+1}\rangle),$$

where h_i a quadratic function parameterized by α_i, defined as follow:

$$h_i(x) = \begin{cases} \frac{C\alpha_i^2}{4} & |x| \leq \alpha_i \\ -\frac{C(|x|-\alpha_i)^2}{2} + \frac{C\alpha_i^2}{4} & \alpha_i \leq |x| \leq \frac{3}{2}\alpha_i \\ \frac{C(|x|-2\alpha_i)^2}{2} & \frac{3}{2}\alpha_i \leq |x| \leq 2\alpha_i \\ 0 & 2\alpha_i \leq |x|. \end{cases}$$

For technical reasons, we define a truncated perturbation distribution $\mathcal{P}$ as follows. Given a sample U from a d-dimensional isotropic Gaussian $N(0, \mathrm{Id}_d)$, we truncate the i-th coordinate of U so that $\tilde{U}_i = \min(U_i, a_i)$, for some fixed $a_i > 0$ that we will specify below, for all $i = 0, 1, \ldots, d-1$. We let $\mathcal{P}$ denote the distribution of $\tilde{U}$.

The proof of Theorem 4.5 is divided into two cases. In the first, we examine the case when the averaged learning rate is $O(T^{-1/2})$.

Lemma B.2. *In the setting of Theorem 4.5, suppose the learning rates satisfy that* $\sum_{i=0}^{T-1} \eta_i \leq \sqrt{\frac{D^2 kT}{2\sigma^2 C}}$, *consider the function* $f(W)$ *constructed in equation* (8), *we have*

$$\min_{1 \leq t \leq T} \mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \geq D\sqrt{\frac{C\sigma^2}{32kT}}.$$

Proof. We start by defining a gradient oracle by choosing the noise vectors $\{\xi_t\}_{t=0}^{T-1}$ to be independent random variables such that

$$\xi_t = \langle \xi_t, e_{t+1} \rangle e_{t+1} \text{ and } |\langle \xi_t, e_{t+1} \rangle| \leq \frac{\sigma}{\sqrt{k}}, \tag{40}$$

where e_{t+1} is a basis vector whose $(t+1)$-th entry is one and otherwise is zero. In other words, only the $(t+1)$-th coordinate of ξ_t is nonzero, otherwise the rest of the vector remains zero. We use $\bar{\xi}_t$ to denote the averaged noise variable as

$$\bar{\xi}_t = \frac{1}{k} \sum_{i=1}^{k} \xi_t^{(i)},$$

where $\xi_t^{(i)}$ is defined following the condition specified in equation (40). Thus, we can also conclude that

$$|\langle \bar{\xi}_t, e_{t+1} \rangle| \leq \frac{\sigma}{\sqrt{k}}.$$

We consider the objective function $f(W) : \mathbb{R}^d \to \mathbb{R}$ defined above (see also equation (8), Section 4), with

$$\alpha_i = \frac{2\eta_i \sigma}{\sqrt{k}}, \text{ for } i = 0, 1, \ldots, T. \tag{41}$$

We will analyze the dynamics of Algorithm 1 with the objective function $f(W)$ and the starting point $W_0 = D\sqrt{G} \cdot e_0$, where $G = \max\left\{C^{-1}, 2\sum_{i=0}^{T-1} \eta_i\right\}$. For the first iteration, we have

$$\begin{aligned} W_1 &= W_0 - \eta_0 \Big(\frac{1}{2} \sum_{i=1}^{k} \big(\nabla f(W_0 + U_0^{(i)}) + \nabla f(W_0 - U_0^{(i)}) \big) + \bar{\xi}_0 \Big) \\ &= (1 - \eta_0 G^{-1}) W_0 - \eta_0 \bar{\xi}_0, \end{aligned}$$

where U is a random draw from the truncated distribution $\mathcal{P}$ with $\langle U, e_i \rangle = \min\{\mathcal{P}_i, a_i\}$ for $a_i = \frac{\eta_{i-1}\sigma}{\sqrt{k}}$. Next, from the construction of h_1, we get

$$\begin{aligned} &\frac{1}{2}\big(\nabla f(W_1 + U) + \nabla f(W_1 - U)\big) \\ &= G^{-1} \langle W_1, e_0 \rangle e_0 + \frac{1}{2}\Big(h_0'\big(\eta_0 \langle \bar{\xi}_0, e_1 \rangle + \langle U, e_1 \rangle\big) e_1 + h_0'\big(\eta_0 \langle \bar{\xi}_0, e_1 \rangle - \langle U, e_1 \rangle\big) e_1 \Big). \end{aligned}$$

Here, using the fact that $\alpha_0 = \frac{2\eta_0 \sigma}{\sqrt{k}}$ from equation (41) above, and the truncation of U, which implies $|\langle U, e_1 \rangle| \leq \frac{\eta_0 \sigma}{\sqrt{k}}$, and $\langle \bar{\xi}_0, e_1 \rangle \leq \frac{\sigma}{\sqrt{k}}$, we obtain

$$\big|\eta_0 \langle \bar{\xi}_0, e_1 \rangle + \langle U, e_1 \rangle\big| \leq \frac{2\eta_0 \sigma}{\sqrt{k}} = \alpha_0, \text{ and similarly } \big|\eta_0 \langle \bar{\xi}_0, e_1 \rangle - \langle U, e_1 \rangle\big| \leq \frac{2\eta_0 \sigma}{\sqrt{k}} = \alpha_0,$$

which implies that

$$h_0'(\eta_0\langle\bar{\xi}_0, e_1\rangle + \langle U, e_1\rangle) = h_0'(\eta_0\langle\bar{\xi}_0, e_1\rangle - \langle U, e_1\rangle) = 0.$$

This is the first update. Then, in the next iteration,

$$\begin{aligned} W_2 &= W_1 - \eta_1\Big(G^{-1}\langle W_1, e_0\rangle + \bar{\xi}_1\Big) \\ &= -(1-\eta_1 G^{-1})(1-\eta_0 G^{-1})W_0 - \eta_0\bar{\xi}_0 - \eta_1\bar{\xi}_1. \end{aligned}$$

Similarly, we use the fact that $\alpha_i = \frac{2\eta_i\sigma}{\sqrt{k}}$ and the fact that $|\langle U, e_{i+1}\rangle| \leq \frac{\eta_i\sigma}{\sqrt{k}}$, which renders the gradient as zero similar to the above reasoning. This holds for any $i = 1, 2, \ldots, T-1$.

At the t-th iteration, suppose we have that

$$W_t = W_0\prod_{i=0}^{t-1}\Big(1-\eta_i G^{-1}\Big) - \sum_{i=0}^{t-1}\eta_i\bar{\xi}_i.$$

Then by induction, at the $(t+1)$-th iteration, we must have

$$\begin{aligned} W_{t+1} &= W_t - \eta_t\Big(G^{-1}\langle W_t, e_0\rangle + \bar{\xi}_t\Big) \\ &= W_0\prod_{i=0}^{t}\Big(1-\eta_i G^{-1}\Big) - \sum_{i=0}^{t}\eta_i\bar{\xi}_i. \end{aligned} \tag{42}$$

Next, from the definition of h_t above, we have that

$$\begin{aligned} F(W_0) - \min_{W\in\mathbb{R}^d} F(W) &= F(W_0) && \text{(the minimum can be attained at zero)} \\ &= \frac{1}{2G}(D\sqrt{G})^2 + \sum_{i=0}^{T-1}\frac{C}{4}\Big(\frac{2\eta_i\sigma}{\sqrt{k}}\Big)^2 && \text{(since } \langle W_0 + U, e_{i+1}\rangle \leq \alpha_i\text{)} \end{aligned}$$

The above must be at most D^2, which implies that we should set the learning rates to satisfy (after some calculation)

$$\frac{1}{T}\Big(\sum_{i=0}^{T-1}\eta_i\Big)^2 \leq \sum_{i=0}^{T-1}\eta_i^2 \leq \frac{kD^2}{2C\sigma^2}. \tag{43}$$

We note that for all $z \in [0,1]$, $1 - \frac{z}{2} \geq \exp(\log\frac{z}{2})$. Thus, applying this to the right-hand side of equation (42), we obtain that for any t,

$$\prod_{i=0}^{t}\Big(1-\eta_i G^{-1}\Big) \geq \frac{1}{2}, \tag{44}$$

where we recall that $G = \max\{C^{-1}, 2\sum_{i=0}^{T-1}\eta_i\}$. Essentially, our calculation so far shows that for all the h_i except h_0, the algorithm has not moved at all from its initialization at W_0 under the above gradient noise. We thus conclude that

$$\begin{aligned} \min_{1\leq i\leq T}\|\nabla F(W_i)\|^2 &= \min_{1\leq i\leq T}\Big(G^{-1}\langle W_0, e_0\rangle\Big)^2 && \text{(by the construction of } F(\cdot)\text{)} \\ &\geq \frac{1}{4}G^{-2}(D\sqrt{G})^2 && \text{(by equations (42) and (44))} \\ &= \frac{D^2}{4}\min\Big\{C, \frac{1}{2\sum_{i=0}^{T-1}\eta_i}\Big\} && \text{(recall the definition of } G \text{ above)} \\ &\geq \frac{D^2}{4}\min\Big\{C, \frac{\sqrt{2C\sigma^2}}{2D\sqrt{kT}}\Big\} && \text{(by equation (43))} \\ &\geq D\sqrt{\frac{C\sigma^2}{32kT}}. \end{aligned}$$

In the first step, we use the fact that $\langle \bar{\xi}_i, e_0 \rangle = 0$, for all $0 = 1, 2, \ldots, T-1$.

Thus, we have proved that equation (9) holds for W_i for any $i = 1, 2, \ldots, T$. The proof of Lemma B.2 is finished. □

Next, let us consider the case of large, fixed learning rates.

Lemma B.3. *In the setting of Theorem 4.5, suppose the learning rates satisfy that* $\sum_{i=0}^{T-1} \eta_i \geq \sqrt{\frac{D^2 kT}{2\sigma^2 C}}$ *and* $\eta_i = \eta$ *for some fixed* $\eta \leq C^{-1}$. *Then, consider the function from equation (8), we have that* $\min_{1 \leq t \leq T} \mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \geq D\sqrt{\frac{C\sigma^2}{32kT}}$.

Proof. We define the functions g, parametrized by a fixed, positive constants $\alpha = \frac{1-\rho^T}{1-\rho} \cdot 2c\eta\sigma$, as follows:

$$g(x) = \begin{cases} -\frac{C}{2}x^2 + \frac{C}{4}\alpha^2 & |x| \leq \frac{\alpha}{2}, \\ \frac{C}{2}(|x| - \alpha)^2 & \frac{\alpha}{2} \leq |x| \leq \alpha, \\ 0 & \alpha \leq |x|. \end{cases}$$

One can verify that g has C-Lipschitz gradient, but g is not twice-differentiable. We also consider a chain-like function:

$$f(W) = g(\langle W, e_0 \rangle) + \sum_{t=0}^{d-1} \frac{C}{2} \langle W, e_{t+1} \rangle^2. \tag{45}$$

From the definition of f, f also has C-Lipschitz gradient. Similar to equation (40), we start by defining an adversarial gradient oracle by choosing the noise vectors $\{\xi_t\}_{t=0}^{T-1}$ to be independent random variables such that

$$\xi_t = \langle \xi_t, e_{t+1} \rangle, \mathbb{E}\left[\langle \xi_t, e_{t+1} \rangle^2\right] = \sigma^2, \text{ and } |\langle \xi_t, e_{t+1} \rangle| \leq c\sigma,$$

where c is a fixed constant. We use $\bar{\xi}_t$ to denote the averaged noise variable as

$$\bar{\xi}_t = \sum_{i=1}^{k} \xi_t^{(i)}.$$

Suppose $\{\xi_t^{(i)}\}_{i=1}^k$ are i.i.d. random variables for any t, we have

$$|\langle \bar{\xi}_t, e_{t+1} \rangle| \leq c\sigma \text{ and } \mathbb{E}\left[\left\|\bar{\xi}_t\right\|^2\right] \leq \frac{\sigma^2}{k}. \tag{46}$$

Next, we analyze the dynamics of Algorithm 1 with the objective function $f(W)$ and the starting point $W_0 = \sum_{i=1}^d \sqrt{\frac{D^2}{Cd}} \cdot e_i$. In this case, by setting $\eta_i = \eta$ for all $i = 0, 1, \ldots, T-1$. Recall that $\eta < C^{-1}$. Denote by $\rho = C\eta$, which is strictly less than one.

Since h_t is an even function, its derivative h_t' is odd. For the first iteration, we have

$$\begin{aligned} W_1 &= W_0 - \eta\Big(\frac{1}{2}\big(\nabla f(W_0 + U) + \nabla f(W_0 - U)\big) + \bar{\xi}_0\Big) \\ &= (1 - C\eta)W_0 - \eta\bar{\xi}_0. \end{aligned}$$

where U is a truncate distribution of $\mathcal{P} \sim N(0, \mathrm{Id}_d)$ with $\langle U, e_0 \rangle = \min\{\mathcal{P}_0, a_0\}$ and $a_0 = c\eta\sigma$.

Using the fact that $\alpha = \frac{1-\rho^T}{1-\rho} \cdot 2c\eta\sigma$, $|\langle U, e_0 \rangle| \leq c\eta\sigma$, and $\langle \bar{\xi}_0, e_0 \rangle \leq c\sigma$, we have

$$g'(\eta\langle \bar{\xi}_0, e_0 \rangle + \langle U, e_0 \rangle) + g'(\eta\langle \bar{\xi}_0, e_0 \rangle - \langle U, e_0 \rangle) = -2C\eta\langle \bar{\xi}_0, e_0 \rangle.$$

Then, in the next iteration,

$$\begin{aligned} W_2 &= W_1 - \eta\Big(C\sum_{i=1}^{d}\langle W_1, e_i\rangle - C\eta\bar{\xi}_0 + \bar{\xi}_1\Big) \\ &= (1-C\eta)^2 W_0 - (1-C\eta)\eta\bar{\xi}_0 - \eta\bar{\xi}_1. \end{aligned}$$

Similarly, we use the fact that $\alpha = \frac{1-\rho^T}{1-\rho}\cdot 2c\eta\sigma$ and the fact that $|\langle U, e_0\rangle| \le c\eta\sigma$, which renders the gradient as $g'(x) = -Cx$, for any $i = 1, 2, \ldots, T-1$.

At the t-th iteration, suppose that

$$W_t = (1-C\eta)^t W_0 - \sum_{i=0}^{t-1}(1-C\eta)^{t-1-i}\eta\bar{\xi}_i.$$

Then by induction, at the $(t+1)$-th iteration, we have

$$\begin{aligned} W_{t+1} &= W_t - \eta\Big(C\sum_{i=1}^{d}\langle W_t, e_i\rangle - C\sum_{i=0}^{t-1}(1-C\eta)^{t-1-i}\eta\bar{\xi}_i + \bar{\xi}_t\Big) \\ &= (1-C\eta)^{t+1}W_0 - \sum_{i=0}^{t}(1-C\eta)^{t-1-i}\eta\bar{\xi}_i. \end{aligned} \tag{47}$$

Next, from the definition of F above, we have that

$$\begin{aligned} F(W_0) - \min_{W\in\mathbb{R}^d} F(W) &= F(W_0) \\ &= \frac{dC}{2}\Big(\sqrt{\frac{D^2}{Cd}}\Big)^2 + \frac{C}{4}\Big(\frac{2(1-\rho^T)c\eta\sigma}{(1-\rho)}\Big)^2, \qquad \text{(since } \langle W_0 + U, e_0\rangle \le \alpha) \end{aligned}$$

which must be at most D^2. Thus, we must have (after some calculation)

$$c^2 \le \frac{D^2(1-\rho)^2}{2\sigma^2\rho^2(1-\rho^T)^2}.$$

We conclude that

$$\begin{aligned} \min_{1\le i\le T}\mathbb{E}\left[\|\nabla F(W_i)\|^2\right] &= \min_{1\le i\le T}\mathbb{E}\left[\sum_{j=1}^{d} C^2\langle W_i, e_j\rangle^2 + C^2\langle W_i, e_0\rangle^2\right] \\ &= \min_{1\le i\le T}\Big(dC^2(1-\rho)^{2t}\Big(\sqrt{\frac{D^2}{Cd}}\Big)^2 + \frac{\sigma^2}{k}\cdot\rho^2\sum_{i=0}^{t}(1-\rho)^{2(t-1-i)}\Big) \\ &\ge \min_{1\le i\le T}\Big(CD^2(1-\rho)^{2t} + \frac{\sigma^2}{k}\frac{\rho}{2-\rho}\big(1-(1-\rho)^{2t}\big)\Big) \\ &\ge \min\Big\{CD^2, \frac{\sigma^2}{k}\frac{\rho}{2-\rho}\Big\} \\ &\ge \frac{\sigma^2}{k}C\sqrt{\frac{kD^2}{2T\sigma^2C}}\frac{1}{2-C\sqrt{\frac{kD^2}{2T\sigma^2C}}} \\ &\ge D\sqrt{\frac{C\sigma^2}{16k\cdot T}}. \qquad \text{(after some calculation)} \end{aligned}$$

Thus, we have proved this lemma. □

Taking both Lemma B.2 and B.3 together, we thus conclude the proof of Theorem 4.5.

B.3 Proof of momentum lower bound

In this section, we prove the following result.

Theorem B.4. *There exists a quadratic function f such that for the iterates $W_1, \ldots, W_T$ generated by equation (10), we must have:* $\min_{1\le t\le T} \mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \ge O\big(D\sqrt{\frac{C\sigma^2}{k\cdot T}}\big)$.

We will focus on a perturbation distribution $\mathcal{P}$ equal to the isotropic Gaussian distribution for this result. In this case, we know that $F(W) = f(W) + d$. For the quadratic function $f(W) = \frac{C}{2}\|W\|^2$, its gradient is clearly C-Lipschitz. We set the initialization $W_0 \in \mathbb{R}^d$ such that

$$F(W_0) - \min_{W\in\mathbb{R}^d} F(W) = D^2.$$

This condition can be met when we set W_0 as a vector whose Euclidean norm is equal to

$$D\sqrt{2\max\Big\{C^{-1}, 2\sum_{i=0}^{T-1}\eta_i\Big\}}.$$

The case when $\mu = 0$. We begin by considering the case when $\mu = 0$. In this case, the update reduces to SGD, and the iterate W_{t+1} evolves as follows:

$$W_{t+1} = \Big(1 - C\eta_t\Big)W_t - \eta_t\bar{\xi}_t, \tag{48}$$

where we denote $\bar{\xi}_t$ as the averaged noise $k^{-1}\sum_{j=1}^k \xi_t^{(j)}$, and the noise perturbation $U_t^{(j)}$ cancelled out between the plus and minus perturbations. The case when $\mu > 0$ builds on this simpler case, as we will describe below.

The key observation is that the gradient noise sequence $\bar{\xi}_1, \bar{\xi}_2, \ldots, \bar{\xi}_T$ forms a martingale sequence:

- For any $i = 1, 2, \ldots, T$, conditioned on the previous random variables $\xi_{i'}^{(j)}$ for any $i' < i$ and any $j = 1, 2, \ldots, k$, the expectation of $\bar{\xi}_i$ is equal to zero.
- In addition, the variance of $\bar{\xi}_i$ is equal to $k^{-1}\sigma^2$, since conditional on the previous random variables, the $\xi_i^{(j)}$s are all independent from each other.

The martingale property allows us to characterize the SGD path of $\|W_t\|^2$, as shown in the following result.

Lemma B.5. *In the setting of Theorem B.4, for any step sizes $\eta_0, \ldots, \eta_{T-1}$ less than C^{-1}, and any $t = 1, \ldots, T$, the expected gradient of W_t, $\mathbb{E}\left[\|\nabla F(W_t)\|^2\right]$, is equal to*

$$2CD^2\prod_{j=0}^{t-1}\big(1 - C\eta_j\big)^2 + \frac{C\sigma^2}{k}\sum_{i=0}^{t-1}\eta_i^2\prod_{j=i+1}^{t-1}\big(1 - C\eta_j\big)^2.$$

Proof. By iterating over equation (48), we can get

$$W_t = W_0\prod_{j=0}^{t-1}\Big(1 - C\eta_j\Big) - \sum_{i=0}^{t-1}\eta_i\bar{\xi}_i\prod_{j=i+1}^{t-1}\Big(1 - C\eta_j\Big).$$

Meanwhile,

$$\nabla F(W_t) = CW_t \Rightarrow \|\nabla F(W_t)\|^2 = C^2\|W_t\|^2.$$

Thus, by squaring the norm of W_t and taking the expectation, we can get

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] = C^2 \|W_0\|^2 \prod_{j=0}^{t-1} \left(1 - C\eta_j\right)^2 + C^2 \sum_{i=0}^{t-1} \mathbb{E}\Big[\Big|\Big|\eta_i \bar{\xi}_i \prod_{j=i+1}^{t-1} \left(1 - C\eta_j\right)\Big|\Big|^2\Big]. \tag{49}$$

Above, we use martingale property a), which says the expectation of $\bar{\xi}_i$ is equal to zero for all i. In addition, based on property b), equation (49) is equal to

$$C^2 \sum_{i=0}^{t-1} \eta_i^2 \left(\prod_{j=i+1}^{t-1} \left(1 - C\eta_j\right)^2 \mathbb{E}\left[\|\bar{\xi}_i\|^2\right] \right) = \frac{C^2\sigma^2}{k} \sum_{i=0}^{t-1} \eta_i^2 \prod_{j=i+1}^{t-1} \left(1 - C\eta_j\right)^2.$$

To see this, based on the martingale property of $\bar{\xi}$ again, the cross terms between $\bar{\xi}_i$ and $\bar{\xi}_j$ for different i, j are equal to zero in expectation:

$$\mathbb{E}\left[\langle \bar{\xi}_i, \bar{\xi}_j \rangle | \bar{\xi}_j\right] = 0, \text{ for all } 1 \leq j < i \leq T.$$

Additionally, the second moment of $\bar{\xi}_i$ satisfies:

$$\mathbb{E}\left[\|\bar{\xi}_i\|^2\right] = \frac{\sigma^2}{k}, \text{ for any } i = 1, \ldots, T.$$

Lastly, let W_0 be a vector such that

$$\|W_0\| = D\sqrt{2C^{-1}} \Rightarrow F(W_0) - \min_{W \in \mathbb{R}^d} F(W) \leq D^2.$$

Setting $\|W_0\| = D\sqrt{2C^{-1}}$ in equation (49) leads to

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] = 2CD^2 \prod_{j=0}^{t-1} \left(1 - C\eta_j\right)^2 + \frac{C^2\sigma^2}{k} \sum_{i=0}^{t-1} \eta_i^2 \prod_{j=i+1}^{t-1} \left(1 - C\eta_j\right)^2.$$

Thus, we conclude the proof of this result. □

We now present the proof for the case when $\sum_{i=0}^{T-1} \eta_i \leq O(\sqrt{T})$. For this result, we will use the following quadratic function:

$$f(W) = \frac{1}{2\kappa} \|W\|^2, \text{ where } \kappa = \max\{C^{-1}, 2\sum_{i=0}^{T-1} \eta_i\}, \tag{50}$$

Lemma B.6. *Consider f given in equation (50) above. For any step sizes $\eta_0, \ldots, \eta_{T-1}$ less than C^{-1}, the following holds for the stochastic objective F:*

$$\min_{1 \leq t \leq T} \mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \geq \frac{D^2}{2\max\{C^{-1}, 2\sum_{i=0}^{T-1} \eta_i\}}.$$

Proof. Clearly, the norm of the gradient of $F(W)$ is equal to

$$\|\nabla F(W)\| = \frac{1}{\kappa}\|W\|. \tag{51}$$

Following the update rule in NSO, similar to equation (48), W_t evolves as follows:

$$W_{t+1} = \left(1 - \frac{\eta_t}{\kappa}\right) W_t - \eta_t \bar{\xi}_t, \tag{52}$$

where $\bar{\xi}_t$ has variance equal to σ^2/k, according to the proof of Lemma B.5. By iterating equation (52) from the initialization, we can get a closed-form equation for $W_t^{(1)}$, for any $t = 1, 2, \ldots, T$:

$$W_t = W_0 \prod_{j=0}^{t-1}\left(1 - \frac{\eta_j}{\kappa}\right) - \sum_{k=0}^{t-1} \eta_k \xi_k \prod_{j=k+1}^{t-1}\left(1 - \frac{\eta_j}{\kappa}\right). \tag{53}$$

Following equation (51), we can show that

$$\|\nabla F(W)\|^2 = \kappa^{-2}\|W_t\|^2.$$

Thus, in expectation,

$$\begin{aligned}
\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] &= \kappa^{-2}\,\mathbb{E}\left[\|W_t\|^2\right] \\
&= \kappa^{-2}\|W_0\|^2 \prod_{j=0}^{t-1}\left(1 - \kappa^{-1}\eta_j\right)^2 + \kappa^{-2}\sum_{i=0}^{t-1}\mathbb{E}\left[\left(\eta_i \bar{\xi}_i \prod_{j=i+1}^{t-1}\left(1 - \kappa^{-1}\eta_j\right)\right)^2\right] \\
&= \kappa^{-2}\|W_0\|^2 \prod_{j=0}^{t-1}\left(1 - \kappa^{-1}\eta_j\right)^2 + \kappa^{-2}\sum_{i=0}^{t-1}\eta_i^2 \prod_{j=i+1}^{t-1}\left(1 - \kappa^{-1}\eta_j\right)^2 \mathbb{E}\left[\|\bar{\xi}_i\|^2\right] \\
&= 2D^2\kappa^{-1}\prod_{j=0}^{t-1}\left(1 - \kappa^{-1}\eta_j\right)^2 + \frac{\sigma^2\kappa^{-2}}{k}\sum_{i=0}^{t-1}\eta_i^2 \prod_{j=i+1}^{t-1}\left(1 - \kappa^{-1}\eta_j\right)^2,
\end{aligned} \tag{54}$$

where we use the definition of initialization W_0 and the variance of $\bar{\xi}_i$ in the last step. In order to tackle equation (54), we note that for all $z \in [0, 1]$,

$$1 - \frac{z}{2} \geq \exp\left(\log\frac{1}{2}\cdot z\right). \tag{55}$$

Hence, applying equation (55) to the right-hand side of equation (54), we obtain that for any $i = 0, 1, \ldots, t-1$,

$$\begin{aligned}
&\prod_{j=i}^{t-1}\left(1 - \frac{\eta_j}{\max\{C^{-1}, 2\sum_{j=i}^{T-1}\eta_i\}}\right) \\
&\geq \exp\left(\log\frac{1}{2}\cdot\sum_{j=i}^{t-1}\frac{\eta_j}{\max\{(2C)^{-1}, \sum_{i=0}^{T-1}\eta_i\}}\right) \geq \frac{1}{2}.
\end{aligned}$$

Thus, equation (54) must be at least

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \geq \frac{2D^2\kappa^{-1}}{4} + \frac{\sigma^2\kappa^{-2}}{k}\sum_{i=0}^{t-1}\frac{\eta_i^2}{4}. \tag{56}$$

The above result holds for any $t = 1, 2, \ldots, T$. Therefore, we conclude that

$$\min_{1\leq t\leq T}\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \geq \frac{D^2}{2\max\{C^{-1}, 2\sum_{i=0}^{T-1}\eta_i\}}.$$

Thus, the proof of Lemma B.6 is finished. □

Next we consider the other case when the learning rates are fixed.

Lemma B.7. *There exists convex quadratic functions f such that for any gradient oracle satisfying Assumption 4.1 and any distribution $\mathcal{P}$ with mean zero, if $\eta_i = \eta < C^{-1}$ for any $i = 1, \ldots, T$, or if $\sum_{i=0}^{T-1} \eta_i \lesssim \sqrt{T}$, then the following must hold:*

$$\min_{1 \le t \le T} \mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \ge D\sqrt{\frac{C\sigma^2}{32k \cdot T}}. \tag{57}$$

Proof. By Lemma B.6, there exists a function such that the left-hand side of equation (57) is at least

$$\frac{D^2}{2\max\{C^{-1}, 2\sum_{i=0}^{T-1}\eta_i\}} \ge \frac{CD^2}{2\max\{1, 2x^{-1}\sqrt{T}\}} = \frac{D^2 x}{4\sqrt{T}}, \tag{58}$$

which holds if $\sum_{i=0}^{T-1} \eta_i \le \sqrt{T}x^{-1}$ for any fixed $x > 0$.

On the other hand, if $\sum_{i=0}^{T-1} \eta_i \ge x^{-1}\sqrt{T}$ and $\eta_i = \eta$ for a fixed η, then $\eta > x^{-1}/\sqrt{T}$. By setting $\eta_i = \eta$ for all i in Lemma B.5, the left-hand side of equation (57) is equal to

$$\min_{1 \le t \le T}\Big(2CD^2(1 - C\eta)^{2t} + \frac{C^2\sigma^2}{k}\sum_{k=0}^{t-1}\eta^2(1 - C\eta)^{2(t-k-1)}\Big).$$

Recall that $\eta < C^{-1}$. Thus, $\rho = C\eta$ must be less than one. With some calculations, we can simplify the above to

$$\begin{aligned}&\min_{1 \le t \le T}\left(2CD^2(1-\rho)^{2t} + \frac{\sigma^2\rho^2}{k}\frac{1-(1-\rho)^{2t}}{1-(1-\rho)^2}\right)\\ &= \min_{1 \le t \le T}\left(\frac{\sigma^2\rho}{k(2-\rho)} + (1-\rho)^{2t}\Big(2CD^2 - \frac{\sigma^2\rho}{k(2-\rho)}\Big)\right).\end{aligned} \tag{59}$$

If $2CD^2 < \frac{\sigma^2\rho}{k(2-\rho)}$, the above is the smallest when $t = 1$. In this case, equation (59) is equal to

$$2CD^2(1-\rho)^2 + \frac{\sigma^2\rho^2}{k} \ge \frac{1}{\frac{1}{2CD^2} + \frac{k}{\sigma^2}} = O(1).$$

If $2CD^2 \ge \frac{\sigma^2\rho}{k(2-\rho)}$, the above is the smallest when $t = T$. In this case, equation (59) is at least

$$\frac{\sigma^2\rho}{k(2-\rho)} \ge \frac{\sigma^2\rho}{2k} \ge \frac{\sigma^2 C x^{-1}}{2k} \cdot \frac{1}{\sqrt{T}}. \tag{60}$$

To conclude the proof, we set x so that the right-hand side of equations (58) and (60) match each other. This leads to

$$x = \sqrt{\frac{2\sigma^2 C}{kD^2}}.$$

Thus, by combining the conclusions from both equations (58) and (60) with this value of x, we finally conclude that if $\sum_{i=0}^{T-1} \eta_i \le \sqrt{T}x^{-1}$, or for all $i = 0, \ldots, T-1$, $\eta_i = \eta < C^{-1}$, then in both cases, there exists a function f such that equation (57) holds. This completes the proof of Lemma B.7. □

The case when $\mu > 0$. In this case, since the update of W_t also depends on the update of the momentum, it becomes significantly more involved. One can verify that the update from step t to step $t+1$ is based on

$$X_u = \begin{bmatrix} 1 - C\eta_t & \mu \\ C\eta_t & \mu \end{bmatrix}. \tag{61}$$

Our analysis examines the eigenvalues of the matrix $X_u X_u^\top$ and the first entry in the corresponding eigenvectors. Particularly, we show that the two entries are bounded away from zero. Then, we apply the Hölder's inequality to reduce the case of $\mu > 0$ to the case of $\mu = 0$, Lemma B.7 in particular.

Proof. First, consider a quadratic function

$$f(W) = \frac{1}{2C} \|W\|^2 .$$

Clearly, $f(W)$ is C-Lipschitz. Further, $F(W) = f(W) + d$, for $\mathcal{P}$ being the isotropic Gaussian. Let W_0 be a vector whose Euclidean norm equals $D\sqrt{2C}$. Thus,

$$F(W_0) - \min_{W \in \mathbb{R}^d} F(W) = D^2.$$

As for the dynamic of momentum SGD, recall that

$$M_{t+1} = \mu M_t - \eta_t G_t \text{ and } \; W_{t+1} = W_t + M_{t+1}.$$

We consider the case where $\eta_t = \eta$ for all steps t. In this case, we can write the above update into a matrix notation as follows:

$$\begin{bmatrix} W_{t+1} \\ M_{t+1} \end{bmatrix} = \begin{bmatrix} 1 - C\eta & \mu \\ -C\eta & \mu \end{bmatrix} \begin{bmatrix} W_t \\ M_t \end{bmatrix} + C\eta \begin{bmatrix} \bar{\xi}_t \\ \bar{\xi}_t \end{bmatrix}.$$

Let $X_\mu = [1 - C\eta, \mu; -C\eta, \mu]$ denote the 2 by 2 matrix (that depends on μ) above. Similar to Lemma B.5, we can apply the above iterative update to obtain the formula for W_{t+1} as:

$$\begin{bmatrix} W_{t+1} \\ M_{t+1} \end{bmatrix} = X_u^t \begin{bmatrix} W_0 \\ M_0 \end{bmatrix} + \sum_{i=0}^{t} C\eta X_u^{t-i} \begin{bmatrix} \bar{\xi}_i \\ \bar{\xi}_i \end{bmatrix}. \tag{62}$$

By multiplying both sides by the vector $e_1 = [1, 0]^\top$, and then taking the Euclidean norm of the vector (notice that this now only evolves that W_{t+1} vector on the left, and the W_t vector on the right), we now obtain that, in expectation over the randomness of the $\bar{\xi}_i$'s, the following holds:

$$\mathbb{E}\left[\|W_{t+1}\|^2\right] = 2CD^2(e_1^\top X_u^t e_1)^2 + \frac{C^2\eta^2\sigma^2}{k} \sum_{i=0}^{t} \left\|e_1^\top X_u^i e\right\|^2 . \tag{63}$$

Above, similar to Lemma B.5, we have set the length of W_0 appropriately, so that its length is equal to $D\sqrt{2C^{-1}}$, which has led to the CD^2 term above. Recall that M_0 is equal to zero in the beginning. To get the first term above, we follow this calculation:

$$\begin{aligned} \left\| e_1^\top X_\mu^t \begin{bmatrix} W_0 \\ M_0 \end{bmatrix} \right\|^2 &= \operatorname{Tr}\left[e_1^\top X_\mu^t \begin{bmatrix} W_0 \\ M_0 \end{bmatrix} \begin{bmatrix} W_0 \\ M_0 \end{bmatrix}^\top {X_\mu^t}^\top e_1 \right] \\ &= \operatorname{Tr}\left[e_1^\top X_\mu^t \begin{bmatrix} CD^2 & 0 \\ 0 & 0 \end{bmatrix} {X_\mu^t}^\top e_1 \right] \\ &= 2CD^2(e_1^\top X_\mu^t e_1)^2. \end{aligned}$$

We use $e = [1, 1]^\top$ to denote the vector of ones. Now, we focus on the 2 by 2 matrix X_u (recall this is the coefficient matrix on the right side of equation (62)). Let its singular values be denoted as λ_1 and λ_2. In addition, to deal with equation (63), let α_1 and α_2 denote the first entry of X_u's left singular vectors, corresponding to a and b, respectively. Thus, we can write

$$(e_1^\top X_\mu^i e)^2 = \alpha_1^2 \lambda_1^{2i} + \alpha_2^2 \lambda_2^{2i}. \tag{64}$$

Now, one can verify that λ_1^2 and λ_2^2 are the roots of the following quadratic equation over x:

$$x^2 - ((1 - C\eta)^2 + (C\eta)^2 + 2\mu^2)x + \mu^2 = 0. \tag{65}$$

This can be checked by first taking X_u times $X_u^\top$, then using the definition of the eigenvalues by calculating the determinant of $X_u X_u^\top - x\,\mathrm{Id} = 0$. Thus, we have that λ_1 and λ_2 are equal to:

$$\lambda_1, \lambda_2 = \frac{(1-C\eta)^2 + (C\eta)^2 + 2\mu^2 \pm \sqrt{((1-C\eta)^2 + (C\eta)^2 + 2\mu^2)^2 - 4\mu^2}}{2}. \tag{66}$$

Now, α_1^2 (and α_2^2, respectively) satisfies that:

$$\alpha_1^2 = \frac{-C\eta(1-C\eta) + \mu^2}{(1-C\eta)^2 + \mu^2 - \lambda_1 + -C\eta(1-C\eta) + \mu^2}. \tag{67}$$

By enumerating the possible values of $C\eta$ between 0 and 1, one can verify that for a fixed value of μ, α_1^2 and α_2^2 are both bounded below from zero. Therefore, we can claim that from equation (64),

$$\alpha_1^2 \lambda_1^{2i} + \alpha_2^2 \lambda_2^{2i} \gtrsim \lambda_1^{2i} + \lambda_2^{2i}. \tag{68}$$

By the Hölder's inequality,

$$(\lambda_1^{2i} + \lambda_2^{2i})^{\frac{1}{2i}} (1+1)^{1-\frac{1}{2i}} \geq \lambda_1 + \lambda_2 = (1-C\eta)^2 + (C\eta)^2 + 2\mu^2 \tag{69}$$

$$\geq (1-C\eta)^2 + (C\eta)^2, \tag{70}$$

which implies that

$$\lambda_1^{2i} + \lambda_2^{2i} \geq \frac{((1-C\eta)^2 + (C\eta)^2)^i}{2^{(2i-1)}}. \tag{71}$$

Now, we consider two cases. If $C\eta < 1/2$, then the above is greater than $(1-C\eta)^{2i}$, which holds for any $i = 0, 1, \ldots, T-1$. By way of reduction, we can follow the proof of Lemma B.7 to complete this proof. If $C\eta > 1/2$, then the above is greater than $(C\eta)^{2i}$. Again by following the proof steps in Lemma B.7, we can show that

$$\min_{t=1}^{T} \mathbb{E}\left[\|W_t\|^2\right] \gtrsim D\sqrt{\frac{C\sigma^2}{k \cdot T}}.$$

This completes the proof of Theorem B.4. □

C Omitted Experiment Details

Comparison of the largest eigenvalue of the loss Hessian. In Figure 5, we report the comparison of the largest eigenvalue, i.e., λ_1, of the Hessian matrix, using the model at the last epoch of fine-tuning, in the same setting as Figure 1. We observe that our algorithm also reduces the λ_1 of the Hessian matrix compared with SAM and SGD.

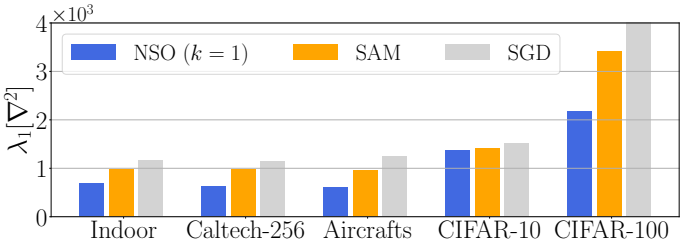


Figure 5: Reporting the λ_1 of the Hessian matrix in the last iteration of fine-tuning ResNet-34 on five data sets, comparing NSO with SAM and SGD. The results are averaged over five random seeds.

Implementation of Hessian measurements in Table 2. Recall that we find that the trace of the Hessian provides an accurate approximation to the gap between the perturbed loss and the trained model loss across several neural networks. These include (1) a two-layer Multi-Layer Perceptron (MLP) trained on the MNIST digit classification data set, (2) a twelve-layer BERT-Base model trained on the MRPC sentence classification data set from the GLUE benchmark, and (3) a two-layer Graph Convolutional Network (GCN) trained on the COLLAB node classification data set from TUDataset.

In more detail, we set both MLP and GCN with a hidden dimension of 128 for model architectures and initialize them randomly. We initialize the BERT model from pretrained BERT-Base-Uncased. We train each model on the provided training set for the training process until the training loss is close to zero. Specifically, we train the MLP, BERT, and GCN models for 30, 10, and 100 epochs. We use the model of the last epoch to measure the error in the approximation.

Implementation. We use the same training hyper-parameters for the experiments in Section 3. These include a learning rate of 0.02, batch size of 32, and training epochs of 30. We reduce the learning rate by 0.1 every 10 epochs. We choose these hyper-parameters based on a grid search on the validation split. The range of hyper-parameters in which we conduct a grid search is as follows:

- Learning rate: 0.05, 0.02, 0.01, 0.005, 0.002, and 0.001;
- Epochs: 10, 20, and 30;
- Batch size: 16, 32, and 64.

Note that we do not use momentum, weight decay, or other advanced techniques in this experiment.

Each baseline method has its own set of hyper-parameters. We also conduct a grid search for the hyper-parameters specifically for each baseline.

- For label smoothing, we choose the weight of the loss calculated from the incorrect labels between 0.1, 0.2, and 0.3.
- For SAM and BSAM, we choose the ℓ_2 norm of the perturbation between 0.01, 0.02, and 0.05.
- For ASAM, we choose the ℓ_2 norm of the perturbation for the rescaled weights between 0.5, 1.0, and 2.0.
- For RSAM, we choose the ℓ_2 norm of the perturbation between 0.01, 0.02, and 0.05 and the standard deviation for sampling perturbation between 0.008, 0.01, and 0.012.