# Noise Stability Optimization For Flat Minima With Tight Rates

**Haotian Ju** JU.H@NORTHEASTERN.EDU
**Dongyue Li** LI.DONGYU@NORTHEASTERN.EDU
**Hongyang R. Zhang** HO.ZHANG@NORTHEASTERN.EDU
*Northeastern University, Boston*

## Abstract

Generalization properties are a central aspect of the design and analysis of learning algorithms. One notion that has been considered in many previous works as leading to good generalization is flat minima, which informally describes a loss surface that is insensitive to noise perturbations. However, the design of efficient algorithms (that are easy to analyze) to find them is relatively under-explored. In this paper, we propose a new algorithm to address this issue, which minimizes a stochastic optimization objective that averages noise perturbations injected into the weights of a function. This algorithm is shown to enjoy both theoretical and empirical advantages compared to existing algorithms involving worst-case perturbations. Theoretically, we show tight convergence rates of our algorithm to find first-order stationary points of the stochastic objective. Empirically, the algorithm induces a penalty on the trace of the Hessian, leading to iterates that are flatter than SGD and other alternatives, with tighter generalization gaps. Altogether, this work contributes a provable and practical algorithm to find flat minima by optimizing the noise stability properties of a function.

## 1. Introduction

The generalization properties of learning algorithms such as stochastic gradient descent (SGD) are central in recent learning theory literature. For instance, implicit regularization refers to a phenomenon where SGD prefers low-norm solutions such as the nuclear norm of a matrix in linear networks [1, 15]. For linearly separable data trained by logistic loss, gradient descent converges to the max-margin solution [16, 22]. As many recent studies have shown, various forms of noise injection and perturbation can also induce implicit bias in the algorithm. For example, injecting label noise (which is also known as label smoothing) induces a nonlinear dynamic at the interpolating regime [4], which corresponds to penalizing large eigenvalues of the Hessian [6].

From a practical standpoint, adding inductive bias to SGD can lead to better empirical performance. A widely used algorithm is sharpness-aware minimization (SAM) [13], based on a (constrained) minimax optimization formulation, which penalizes the worst-case perturbations. However, this minimax optimization problem is challenging to analyze because of the convex-concave constrained formulation, whose computational complexity is NP-hard in the worst case [7]. The algorithm by Foret et al. [13] simplifies the inner maximization by taking one gradient step along the sign of the gradient before computing the gradient. The dynamic of this algorithm is analyzed in Bartlett et al. [3], showing that the algorithm induces an implicit bias on the largest eigenvalue of the Hessian. Despite a strong interest in characterizing the inductive bias and developing learning algorithms that yield generalizable solutions, we seem to have more questions than answers. For

© H. Ju, D. Li & H.R. Zhang.

example, [23] performed various analyses of SAM, including showing the convergence of SAM in the convex quadratic case and the implicit bias induced by a stochastic version of SAM.

The goal of this work is to present a new perspective on finding generalizable solutions by designing a new algorithm that minimizes the average weight perturbations injected into a function. Concretely, let $f : \mathbb{R}^d \to \mathbb{R}$ be a real-valued, nonconvex function. Let $\mathcal{P}$ denote a $d$-dimensional distribution which is symmetric at zero. We consider the following optimization problem, which perturbs the weight of $f$ with a random sample of $\mathcal{P}$:

$$\min_{W \in \mathbb{R}^d} F(W) := \mathop{\mathbb{E}}_{U \sim \mathcal{P}} \left[ f(W + U) \right]. \tag{1}$$

For small enough perturbations, $F(W)$ is approximately equal to $f(W)$ plus a penalty on the Hessian trace of $f(W)$. The difference between the perturbed function and the original $f$ measures *noise stability* properties of a function around a local neighborhood [18], leading algorithms to converge to "flat" regions where the loss surface is less sensitive to perturbations.

The problem of showing explicit convergence rates and algorithmic complexities for minimizing (1) is largely under-explored. The gradient of $F$ cannot be directly computed because it is the average of an infinite population of gradients of $f$. This can be estimated with zeroth-order optimization techniques [9, 12], but it incurs a dimension factor of $\sqrt{d}$. A recent paper by [21] analyzes the regularization effect inducted by noise injection for one-hidden-layer ReLU and deep linear networks. However, non-asymptotic optimization convergence results are still not known. The main contribution of this paper is to show matching upper and lower bounds for minimizing the stochastic objective (1) when $f$ is smooth. Along the way, we design a new algorithm that leverages the symmetry of $\mathcal{P}$ and injects multiple perturbations (for better empirical performance); See Alg. 1.

Our main result shows that this algorithm convergences to a first-order stationary point (of $F$) whose gradient norm squared is at most $O((k \cdot T)^{-1/2})$ after $T$ steps. This is tight, as we construct lower bound instances (both convex and nonconvex) for which the iterates found by SGD (and its variants, including momentum and adaptive learning rates) will have gradients whose norm matches this rate up to constants. From a technical standpoint, we provide a fine-grained characterization of the SGD noise for a convex quadratic function, leveraging a martingale property of the noise sequence. For the nonconvex case, we build on a chain-like construction by [8] while incorporating noise perturbations in the analysis of the lower bound.

We conduct experiments to compare the trace of the Hessian and its largest eigenvalue during the iterations of our algorithm (NSO), SGD, SAM, and SGD with noise injection in the weights. Curiously, across various datasets and neural network architectures, our algorithm finds solutions whose Hessian trace and top eigenvalue are much smaller than all these alternative algorithms. We hypothesize that this is because problem (1) explicitly penalizes the Hessian trace (when the perturbations are small). While our analysis does not formally explain these empirical observations, we believe they justify the design rationale of our algorithm, showing it enjoys practical performance (see Sec. 4 for additional results on test performance). Future works can possibly analyze the solutions preferred by noise-injected SGD in neural nets (cf. literature on "Gaussian smoothing" [19, 20]), or design faster methods for finding local minima of the stochastic objective [5, 11].

**Notations:** 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 $x$. We use the notation $f(x) \lesssim g(x)$ to indicate that $f(x) = O(g(x))$. Let $\mathbf{H}[f(W)]$ denote a $d$ by $d$ matrix whose $(i, j)$-th entry is equal to the twice-derivative of $f(W)$ over the $i$-th and $j$-th entry of $W$. Let $N(0, \mathrm{Id}_d)$ denote the $d$-dimensional isotropic Gaussian.

**Algorithm 1 Noise Stability Optimization (NSO)**

**Input**: Initialization $W_0 \in \mathbb{R}^d$, a gradient estimator $g_z$, a $d$-dimensional distribution $\mathcal{P}$
**Parameters:** Number of epochs $T$, learning rates $\eta_0, \eta_1, \ldots, \eta_{T-1}$, number of perturbations $k$
**Output**: $W_t$

1: **for** $i = 0, 1, \ldots, T-1$ **do**
2: Sample $U_i^{(1)}, U_i^{(2)}, \ldots, U_i^{(k)}$ independently from $\mathcal{P}$
3: Update iterates (with random seeds $z_i^{(1)}, z_i^{(2)}, \ldots, z_i^{(k)}$) according to

$$W_{i+1} = W_i - \frac{\eta_i}{2k} \sum_{j=1}^{k} \left( g_{z_i^{(j)}}\left(W_i + U_i^{(j)}\right) + g_{z_i^{(j)}}\left(W_i - U_i^{(j)}\right)\right) \tag{2}$$

4: **end for**
5: Let $t$ be chosen uniformly at random from $1, 2, \ldots, T$

## 2. Algorithm

This section describes the design of our algorithm based on two rationales: (1) leveraging the symmetry of the perturbation distribution to zero out the first-order term in Taylor's expansion. (2) injecting multiple perturbations to reduce the variance of the stochastic gradient. We state a standard assumption regarding access to the gradient.

**Assumption 1** *Given a random seed $z$, let $g_z : \mathbb{R}^d \to \mathbb{R}^d$ be a (continuous) function providing an unbiased gradient estimate. For any $W \in \mathbb{R}^d$, $g_z$ satisfies the following*

$$\mathop{\mathbb{E}}_{z}\left[g_z(W)\right] = \nabla f(W) \text{ and } \mathop{\mathbb{E}}_{z}\left[\|g_z(W) - \nabla f(W)\|^2\right] \le \sigma^2. \tag{3}$$

Algorithm 1 provides the entire procedure. This algorithm has an implicit bias towards minimizers with low Hessian trace. This can be seen by Taylor's expansion of $F(W)$ (cf. Eq. (1)):

$$\mathop{\mathbb{E}}_{U\sim\mathcal{P}}\left[f(W) + U^\top \nabla f(W) + \frac{1}{2}\langle UU^\top, \mathbf{H}[f(W)]\rangle + \epsilon_U\right] = f(W) + \frac{\sigma^2}{2}\operatorname{Tr}\left[\mathbf{H}[f(W)]\right] + \mathbb{E}\left[\epsilon_U\right], \tag{4}$$

where $\mathcal{P}$ denotes a $d$-dimensional distribution $N(0, \sigma^2 \operatorname{Id}_d)$ and $\epsilon_U$ denotes expansion error terms whose order is $o(\sigma^2)$. Provided with a small enough $\epsilon_U$, $F$ is approximately equal to $f$ plus a penalty on the Hessian trace $\operatorname{Tr}[\mathbf{H}[f]]$.

Based on PAC-Bayesian bounds, a lower Hessian trace can imply a tighter generalization gap. Suppose $f$ is an $l$-layer network, whose $i$-th layer is denoted as $v_i$, for $i = 1, 2, \ldots, l$. Let $\mathbf{H}_i$ be the (loss) Hessian matrix with respect to $v_i$. Based on a linear PAC-Bayes bound (e.g., Theorem 2.1, [10]), the generalization gap between the population and empirical risks can be bounded by a data-dependent bound via the Hessian trace as $\sum_{i=1}^{l} \sqrt{\|v_i\|^2 \operatorname{Tr}\left[\mathbf{H}_i\right]/n}$ (see Theorem 2.1 of Ju et al. [17]). Thus, lowering the Hessian trace leads to a tighter generalization bound.

We experiment with common image and text classification tasks to compare the Hessian trace of the iterates of NSO (Algorithm 1) and SGD. The results are shown in Figure 1, suggesting that NSO reduces the Hessian trace (of the local minimizer in the last epoch) by 60% more than SGD. Interestingly, the generalization gap incurred by NSO is again lower than SGD by over 20%. We will provide more experiments and details later in Sec. 4.

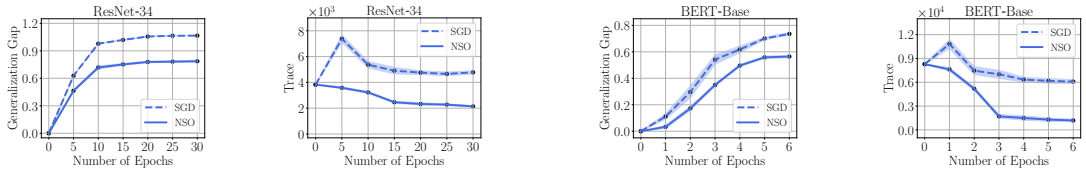


Figure 1: Training curves comparing NSO and SGD; NSO adds a penalty that biases to minimizers with lower Hessian trace compared to SGD and incurs a lower generalization gap.

## 3. Main Results

Our first result gives a fine-grained convergence analysis of the noise stability optimization algorithm. Since finding global minimizers of an arbitrary nonconvex function takes exponential time in the worst case, we focus on finding approximate first-order stationary points in this paper. We assume that $f$ is smooth, whose gradient is Lipschitz-continuous, which implies $F$ is also smooth.

**Theorem 2 (Abridged version of Theorem 5, Sec. A)** *Suppose Assumptions 1 and 3 hold. Let $\mathcal{P}$ be a distribution that is symmetric at zero. There exists a fixed learning rate $\eta < C_1^{-1}$ such that if we run Algorithm 1 with $\eta_i = \eta$ for all $i$, arbitrary number of perturbations $k$, for $T \gtrsim \frac{k}{\sigma^2}$ steps, the algorithm returns $W_t$ such that*

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \lesssim \sqrt{\frac{C_1^3 + C_1\sigma^2}{k \cdot T}}, \tag{5}$$

*in expectation over the randomness of $t$, perturbations, and gradient queries.*

Recall that Algorithm 1 involves averaging the gradients at $W+U$ and $W-U$ for $k$ perturbations $U$. This is equivalent to applying SGD to the following symmetrized loss:

$$F(W) = \mathop{\mathbb{E}}_{U\sim\mathcal{P}}\left[\frac{1}{2}\Big(f(W+U) + f(W-U)\Big)\right]. \tag{6}$$

We provide a motivating example to illustrate the rationale behind the symmetrized gradient.

**Example 1** *Let $u$ be a Gaussian random variable with mean zero and variance $\sigma^2$. Let $f(w) = \frac{1}{4}w^4$. Thus, by definition from Eq. 1, $F(w) = \frac{1}{4}\mathbb{E}_u[(w+u)^4]$. Let us compare the variance of*

$$T = \frac{1}{2}(\nabla f(w+u) + \nabla f(w-u)) \text{ and } G = \frac{1}{2}(\nabla f(w+u) + \nabla f(w+u')),$$

*where $u'$ is another independent sample from the same distribution as $u$. Notice that both $T$ and $G$ require two gradient queries; Interestingly, $T$ has a lower variance than $G$:*

$$\text{var}[T] = 18w^2\sigma^4 < \text{var}[G] = 18w^2\sigma^4 + 4.5w^4\sigma^2 + 7.5\sigma^6. \tag{7}$$

*See Sec. 4 for a comparison between these two gradients in more realistic settings.*

Our proof generalizes this example to characterize the variance of the stochastic gradients of NSO.

Next, we complement the upper bound with a matching lower bound. We give two examples, one on a convex quadratic function, and another on a nonconvex, piecewise quadratic function. For both examples, we show that SGD (and its extensions to momentum and adaptive learning rates), after running $T$ steps, all its iterates have a gradient norm that matches Eq. (10) up constants. Our construction is inspired by the SGD lower bound of [8]. Our analysis gives a fine-grained characterization of the SGD noise. Moreover, the analysis applies to any gradient oracle (with mean zero and bounded variance), which may be of independent interest. For details, see Sec. B.

## 4. Numerical Results

This section conducts experiments to evaluate the empirical benefits of our algorithm. We evaluate Algorithm 1 by fine-tuning pretrained neural nets on image classification datasets, including online medical image classification datasets containing eye fundus images for diabetic retinopathy classification, and several natural image classification datasets, including Aircrafts, Birds, and Indoor. We compare our algorithm with stochastic gradient descent (SGD) and sharpness-aware minimization (SAM). We sample perturbations from isotropic Gaussian distribution $N(0, \sigma^2 \operatorname{Id}_d)$ and tune $\sigma$ among $0.008, 0.01$, and $0.012$. To evaluate the trace and eigenvalues of the loss Hessian matrix, we use Hessian vector products. Other details about the implementation are provided in App. C.

First, we justify both rationales of our design in Sec. 2. The first compares NSO with weight-perturbed SGD using two independently sampled random gradient estimates (cf. Example 1). We validate that NSO can reduce the largest eigenvalue ($\lambda_1$) and the Hessian trace by 8.6% more. The second studies the effect of increasing $k$, as using larger $k$ in NSO would lead to a better estimation of the gradient. We vary $k = 1, 2, 3$ on both datasets. The results are shown in Table 1 below.

Table 1: We describe several ablation studies to understand the relationship between different algorithms. (1) With the same number of gradient queries, NSO can converge to minimizers with lower $\lambda_1$ than weight-perturbed SGD, and achieve better test accuracy. (2) As we add more perturbations, the $\lambda_1$ of Hessian decreases further.

| Dataset: Aircrafts | Test accuracy (↑) | $\lambda_1[\mathbf{H}]$ (↓) | Dataset: Indoor | Test accuracy (↑) | $\lambda_1[\mathbf{H}]$ (↓) |
|---|---|---|---|---|---|
| WP-SGD ($k = 2$) | 60.63±0.38 | 675±58 | WP-SGD ($k = 2$) | 76.49±0.40 | 748±53 |
| NSO ($k = 1$) | **62.31**±0.36 | **612**±44 | NSO ($k = 1$) | **77.39**±0.55 | **688**±58 |
| WP-SGD ($k = 4$) | 60.40±0.39 | 597±47 | WP-SGD ($k = 4$) | 77.11±0.31 | 688±44 |
| NSO ($k = 2$) | **63.46**±0.43 | **567**±63 | NSO ($k = 2$) | **77.44**±0.30 | **652**±56 |
| WP-SGD ($k = 6$) | 60.79±0.69 | 685±63 | WP-SGD ($k = 6$) | 77.34±0.44 | 658±37 |
| NSO ($k = 3$) | **63.64**±0.68 | **550±39** | NSO ($k = 3$) | **77.94**±0.64 | **630**±74 |

Next, we report results comparing both the Hessian trace and $\lambda_1$ of iterates at the final epoch. Curiously, NSO achieves notably smaller $\lambda_1$ and Hessian trace. Figure 2 visualizes this comparison. Due to space limit, the detailed results (including a comparison of empirical test performance) are reported in Table 3, App. C.

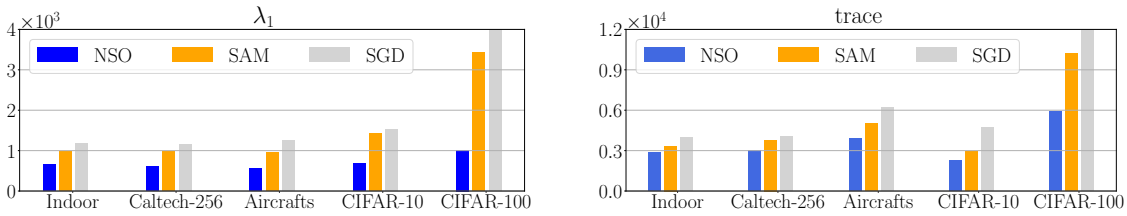


Figure 2: Comparing the loss Hessian's largest eigenvalue and trace between NSO, SAM, and SGD, conducted on ResNet-34, averaged over three random seeds. Interestingly, NSO finds minimizers with lower Hessian trace than SAM. We hypothesize this is because the algorithm minimizes an objective that involves an explicit penalty on the Hessian trace. Formally explaining this behavior is an interesting question for future work.

## References


[1] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Francis Bach. Learning theory from first principles. *Online version*, 2021.

[3] Peter L Bartlett, Philip M Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *arXiv:2210.01513*, 2022.

[4] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *COLT*, 2020.

[5] Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Faster perturbed stochastic gradient methods for finding local minima. In *International Conference on Algorithmic Learning Theory*, pages 176–204. PMLR, 2022.

[6] Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. *NeurIPS*, 2021.

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

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

[9] 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.

[10] 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*, pages 604–612. PMLR, 2021.

[11] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *NeurIPS*, 2018.

[12] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

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

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

[15] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

[16] Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

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

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

[19] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

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

[21] Antonio Orvieto, Anant Raj, Hans Kersting, and Francis Bach. Explicit regularization in overparametrized models via noise injection. *AISTATS*, 2023.

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

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

## Appendix A. Upper Bound

We state the following assumptions.

**Assumption 3** *Let $C_1, C_2, D$ be fixed, positive constants. First, the initialization $W_0 \in \mathbb{R}^d$ satisfies $F(W_0) - \min_{W \in \mathbb{R}^d} F(W) \leq D^2$. Second, the gradient of $f(W)$ exists and is $C_1$-Lipschitz continuous: For any $W_1 \in \mathbb{R}^d$ and $W_2 \in \mathbb{R}^d$, the following holds*

$$\|\nabla f(W_2) - \nabla f(W_1)\| \leq C_1 \|W_2 - W_1\| . \tag{8}$$

A simple corollary of the above is that $\nabla F(W)$ is also Lipschitz-continuous.

**Proposition 4** *Suppose condition (8) holds. Then, $\nabla F(W)$ is also $C_1$-Lipschitz-continuous.*

To see this, we apply Eq. (8) inside the expectation of $F(W)$. For any $W_1, W_2 \in \mathbb{R}^d$, by definition,

$$\begin{aligned}
\|\nabla F(W_1) - \nabla F(W_2)\| &= \left\| \nabla \mathop{\mathbb{E}}_{U \sim \mathcal{P}} [f(W_1 + U)] - \nabla \mathop{\mathbb{E}}_{U \sim \mathcal{P}} [f(W_2 + U)] \right\| \\
&= \left\| \mathop{\mathbb{E}}_{U \sim \mathcal{P}} [\nabla f(W_1 + U) - \nabla f(W_2 + U)] \right\| \\
&\leq \mathop{\mathbb{E}}_{U \sim \mathcal{P}} [\|\nabla f(W_1 + U) - \nabla f(W_2 + U)\|] \leq C_1 \|W_1 - W_2\| .
\end{aligned}$$

In the event that $f(W)$ is twice-differentiable and the Hessian matrix of $f(W)$ is $C_2$-Lipschitz continuous, then the following holds:

$$\left\| \mathbf{H}\big[f(W_2)\big] - \mathbf{H}\big[f(W_1)\big] \right\|_2 \leq C_2 \|W_2 - W_1\| , \text{ for any } W_1, W_2 \in \mathbb{R}^d. \tag{9}$$

For a random variable $U \sim \mathcal{P}$, let the second moment $\mathbb{E}[\|U\|^2]$ be denoted as $H(\mathcal{P})$ and the fourth moment $\mathbb{E}[\|U\|^4]$ be $G(\mathcal{P})$. Now we can state our first result. There are two cases, depending on whether or not the Hessian exists.

**Theorem 5** *Suppose Assumptions 1 and 3 hold. Let $\mathcal{P}$ be a distribution that is symmetric at zero. There exists a fixed learning rate $\eta < C_1^{-1}$ such that if we run Algorithm 1 with $\eta_i = \eta$ for all $i$, arbitrary number of perturbations $k$, for $T \gtrsim \frac{k}{\sigma^2}$ steps, the algorithm returns $W_t$ such that*

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \leq 4D\sqrt{\frac{C_1^3 H(\mathcal{P}) + C_1\sigma^2}{k \cdot T}}, \tag{10}$$

*in expectation over the randomness of $t$, perturbations, and gradient queries. Additionally, when $f(\cdot)$ is twice-differentiable and its Hessian is $C_2$-Lipschitz-continuous, then the following is also true*

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \leq 4D\sqrt{\frac{C_1 C_2^2 G(\mathcal{P}) + C_1\sigma^2}{k \cdot T}}. \tag{11}$$

The only difference in (10) and (11) is their dependence on $\mathcal{P}$. Let us consider $\mathcal{P} = N(0, \sigma^2 \operatorname{Id}_d)$ for example. For small perturbations, the fourth moment $G(\mathcal{P}) = (d^2 + 2d)\sigma^4$ is smaller than the second moment $H(\mathcal{P}) = d\sigma^2$. More precisely, when $\sigma^2 \leq C_1^2 C_2^{-2}(d+2)^{-1}$, Eq. (11) gives a tighter bound than (10).

### A.1. Proof Overview

Next, we characterize the variance of the stochastic gradient. For $i = 0, \dots, T-1$, denote

$$\delta_i := \frac{1}{2k} \sum_{j=1}^{k} \Big( \nabla f\big(W_i + U_i^{(j)}\big) + \nabla f\big(W_i - U_i^{(j)}\big) \Big) - \nabla F(W_i) \tag{12}$$

$$\xi_i := \frac{1}{2k} \sum_{j=1}^{k} \Big( g_{z_i^{(j)}}\big(W_i + U_i^{(j)}\big) - \nabla f\big(W_i + U_i^{(j)}\big) + g_{z_i^{(j)}}\big(W_i - U_i^{(j)}\big) - \nabla f\big(W_i - U_i^{(j)}\big) \Big). \tag{13}$$

It is not hard to see that the mean of $\delta_i$ and $\xi_i$ are both zero. The former is by the symmetry of $\mathcal{P}$ (cf. Eq. (6)). The latter is because $g_z$ is unbiased (cf. Eq. (3)). Next, we give the second moment of $\delta_i$ and $\xi_i$. The key is that both of their variances scale as $\frac{1}{k}$, where $k$ is the number of noise injections.

**Lemma 6** *For any $i = 1, \dots, T$, suppose $\nabla f(W)$ is $C_1$-Lipschitz, then*

$$\mathbb{E}\left[\|\xi_i\|^2\right] \le \frac{\sigma^2}{k} \;\; \textit{and} \;\; \mathbb{E}\left[\|\delta_i\|^2\right] \le \frac{C_1^2 H(\mathcal{P})}{k}. \tag{14}$$

*Additionally, suppose $\mathbf{H}[f(W)]$ is $C_2$-Lipschitz, then*

$$\mathbb{E}\left[\|\delta_i\|^2\right] \le \frac{3C_2^2 G(\mathcal{P})}{8k}. \tag{15}$$

The proof of Lemma 6, based on Proposition 4, can be found in App. D.1. These results are needed in the analysis of the gradient dynamic. In particular, the following result is based on results in [14] (cf. App. D.2).

**Lemma 7** *In the setting of Theorem 5, for any $\eta_0, \cdots, \eta_{T-1}$ less than ${C_1}^{-1}$ and a random variable according to a distribution $\Pr[t = j] = \frac{\eta_j}{\sum_{i=0}^{T-1} \eta_i}$, for any $j = 0, \dots, T-1$, the following holds:*

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \le \frac{2C_1}{\sum_{i=0}^{T-1} \eta_i} D^2 + \frac{C_1 \sum_{i=0}^{T-1} \eta_i^2 \big( \mathbb{E}\left[\|\delta_i\|^2\right] + \mathbb{E}\left[\|\xi_i\|^2\right] \big)}{\sum_{i=0}^{T-1} \eta_i}. \tag{16}$$

**Proof of Theorem 5** Let the step sizes be equal to a fixed $\eta$ for all epochs. Thus, Eq. (16) becomes

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \le \frac{2}{T\eta} D^2 + \frac{C_1 \eta}{T} \sum_{i=0}^{T-1} \Big( \mathbb{E}\left[\|\delta_i\|^2\right] + \mathbb{E}\left[\|\xi_i\|^2\right] \Big). \tag{17}$$

Minimizing over $\eta$ above leads us to the following upper bound on the right-hand side of (17)

$$2T^{-1}\sqrt{2C_1} \cdot D \sqrt{\sum_{i=0}^{T-1} \Big( \mathbb{E}\left[\|\delta_i\|^2\right] + \mathbb{E}\left[\|\xi_i\|^2\right] \Big)}.$$

The proof of Eqs. (10) and (11) follows by applying Lemma 6 to the above. Recall that the step size $\eta$ must be less than $\eta \le {C_1}^{-1}$. Thus, we must have the following

$$\eta = \frac{D}{\sqrt{\sum_{i=0}^{T-1} \Big( \mathbb{E}\left[\|\delta_i\|^2\right] + \mathbb{E}\left[\|\xi_i\|^2\right] \Big)}} \sqrt{\frac{2}{C_1}} \le \frac{1}{C_1} \Rightarrow \frac{2C_1 D^2 k}{C_1^2 H(\mathcal{P}) + \sigma^2} \le T,$$

by applying Eq. (14) to the above. For Eq. (11), the condition is similar and is omitted. In either case, as long as $T \gtrsim \frac{k}{\sigma^2}$, then Eqs. (10) and (11) must hold. ■

## Appendix B. Lower Bound

### B.1. A Convex Example

We focus on iterative algorithms in the form of update (2) in Algorithm 1. We consider a quadratic function

$$f(W) = \frac{C_1}{2} \|W\|^2 . \tag{18}$$

The argument holds for any distribution $\mathcal{P}$ with mean zero. The stochastic objective $F(W)$ is equal to $f(W) + C_1 \cdot H(\mathcal{P})/2$. The SGD path of Algorithm 1 evolves as follows:

$$\begin{aligned} W_{t+1} &= W_t - \eta_t \cdot \frac{1}{2k} \sum_{j=1}^{k} \left( C_1 \Big( W_t + U_t^{(j)} \Big) + C_1 \Big( W_t - U_t^{(j)} \Big) \right) - \eta_t \bar{\xi}_t, \\ &= \Big( 1 - C_1 \eta_t \Big) W_t - \eta_t \bar{\xi}_t, \end{aligned} \tag{19}$$

where we denote $\bar{\xi}_t$ as the averaged noise $\frac{1}{k}\sum_{j=1}^{k} \xi_t^{(j)}$. The key observation is that the gradient noise sequence $\bar{\xi}_1, \bar{\xi}_2, \ldots, \bar{\xi}_T$ forms a martingale sequence:

a) 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.

b) The variance of $\bar{\xi}_i$ is equal to $\frac{\sigma^2}{k}$, since conditional on the previous random variables, the $\xi_i^{(j)}$s are all independent from each other.

This martingale property lets us characterize the SGD path of $W_t$, as shown in the following result.

**Lemma 8** *Consider $f$ specified in Eq. (18). For any step sizes $\eta_0, \ldots, \eta_{T-1}$ less than $C_1^{-1}$, we have*

$$\min_{1 \le t \le T} \mathbb{E} \left[ \|\nabla F(W_t)\|^2 \right] = \min_{1 \le t \le T} \left( 2C_1 D^2 \prod_{j=0}^{t-1} \left(1 - C_1 \eta_j\right)^2 + \frac{C_1 \sigma^2}{k} \sum_{i=0}^{t-1} \eta_i^2 \prod_{j=i+1}^{t-1} \left(1 - C_1 \eta_j\right)^2 \right). \tag{20}$$

**Proof** By iterating over Eq. (19), we can get

$$W_t = W_0 \prod_{j=0}^{t-1} \Big(1 - C_1 \eta_j\Big) - \sum_{i=0}^{t-1} \eta_i \bar{\xi}_i \prod_{j=i+1}^{t-1} \Big(1 - C_1 \eta_j\Big). \tag{21}$$

Meanwhile,

$$\nabla F(W_t) = \nabla f(W_t) = C_1 W_t \Rightarrow \|\nabla F(W_t)\|^2 = C_1^2 \|W_t\|^2 .$$

Thus, following Eq. (21), we can get

$$\mathbb{E} \left[ \|\nabla F(W_t)\|^2 \right] = C_1^2 \|W_0\|^2 \prod_{j=0}^{t-1} \Big(1 - C_1 \eta_j\Big)^2 + C_1^2 \sum_{i=0}^{t-1} \mathbb{E} \left[ \left\| \eta_i \bar{\xi}_i \prod_{j=i+1}^{t-1} \Big(1 - C_1 \eta_j\Big) \right\|^2 \right] \tag{22}$$

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), the second term in the right of Eq. (22) is equal to

$$C_1^2 \sum_{i=0}^{t-1} \eta_i^2 \left( \prod_{j=i+1}^{t-1} \Big(1 - C_1 \eta_j\Big)^2 \mathbb{E} \left[ \|\bar{\xi}_i\|^2 \right] \right) = \frac{C_1^2 \sigma^2}{k} \sum_{i=0}^{t-1} \eta_i^2 \prod_{j=i+1}^{t-1} \Big(1 - C_1 \eta_j\Big)^2 .$$

To see this, based on the martingale property of $\bar{\xi}$:

- 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 $\bar{\xi}_i$ is equal to $\frac{\sigma^2}{k}$:

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

Lastly, let $W_0$ be a vector such that

$$\|W_0\| = D\sqrt{2C_1^{-1}} \Rightarrow F(W_0) - \min_{W \in \mathbb{R}^d} F(W) \leq D^2.$$

Inserting the above value of $W_0$ to Eq. (22) leads to

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] = 2C_1 D^2 \prod_{j=0}^{t-1}\left(1 - C_1\eta_j\right)^2 + \frac{C_1^2\sigma^2}{k}\sum_{i=0}^{t-1}\eta_i^2 \prod_{j=i+1}^{t-1}\left(1 - C_1\eta_j\right)^2 \tag{23}$$

By taking the minimum over $t = 1, \ldots, T$, we conclude that Eq. (20) holds. ■

This result applies to arbitrary learning rates. Next, we show a simpler statement for a slightly different quadratic function

$$f(W) = \frac{1}{2\kappa}\|W\|^2, \text{ where } \kappa = \max\{C_1^{-1}, 2\sum_{i=0}^{T-1}\eta_i\}, \tag{24}$$

which will be easier to work with for small learning rates, as we shall see next.

**Lemma 9** *Consider $f$ given in Eq. (24) above. For any step sizes $\eta_0, \ldots, \eta_{T-1}$ less than $C_1^{-1}$, the following holds for the stochastic objective $F$:*

$$\min_{1 \leq t \leq T} \mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \geq \frac{D^2}{2\max\{C_1^{-1}, 2\sum_{i=0}^{T-1}\eta_i\}}. \tag{25}$$

The proof of Eq. 25 is similar to Lemma 8; See details in App. D.3. Now we are ready to state the main result of this section.

**Theorem 10** *There exists convex quadratic functions $f$ such that for any gradient oracle satisfying Assumption 1 and any distribution $\mathcal{P}$ with mean zero, if $\eta_i = \eta < C_1^{-1}$ for any $i = 1, \ldots, T$, or if $\sum_{i=0}^{T-1}\eta_i \lesssim \sqrt{T}$, then the following must hold:*

$$\min_{1 \leq t \leq T} \mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \geq D\sqrt{\frac{C_1\sigma^2}{32k \cdot T}}. \tag{26}$$

The result of Eq. (26) matches the upper bound for twice-differentiable functions in Eq. (11). More precisely, for small perturbations when the moments $H(\mathcal{P})$ and $G(\mathcal{P})$ are constants, e.g., when $\mathcal{P} = N(0, d^{-1}\,\mathrm{Id}_d)$, the lower bound matches the upper bound up to constants.

#### B.1.1. PROOF OF CONVEX LOWER BOUND

By Lemma 9, there exists a function such that the left-hand side of Eq. (26) is at least

$$\frac{D^2}{2\max\{C_1^{-1}, 2\sum_{i=0}^{T-1}\eta_i\}} \geq \frac{C_1 D^2}{2\max\{1, 2x^{-1}\sqrt{T}\}} = \frac{D^2 x}{4\sqrt{T}}, \tag{27}$$

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

On the other hand, if $\sum_{i=0}^{T-1}\eta_i \geq x^{-1}\sqrt{T}$ and $\eta_i = \eta$ for a fixed $\eta$, then $\eta > x^{-1}/\sqrt{T}$. By setting $\eta_i = \eta$ for all $i$ in Lemma 8, the left-hand side of Eq. (26) is equal to

$$\min_{1\leq t\leq T}\left(2C_1D^2(1-C_1\eta)^{2t} + \frac{C_1^2\sigma^2}{k}\sum_{k=0}^{t-1}\eta^2(1-C_1\eta)^{2(t-k-1)}\right).$$

Recall that $\eta < C_1^{-1}$. Thus, $\rho = C_1\eta$ must be strictly less than one. With some calculation, we can simplify the above to

$$\begin{aligned}&\min_{1\leq t\leq T}\left(2C_1D^2(1-\rho)^{2t} + \frac{\sigma^2\rho^2}{k}\frac{1-(1-\rho)^{2t}}{1-(1-\rho)^2}\right)\\ =&\min_{1\leq t\leq T}\left(\frac{\sigma^2\rho}{k(2-\rho)} + (1-\rho)^{2t}\Big(2C_1D^2 - \frac{\sigma^2\rho}{k(2-\rho)}\Big)\right).\end{aligned} \tag{28}$$

If $2C_1D^2 < \frac{\sigma^2\rho}{k(2-\rho)}$, the above is the smallest when $t = 1$. In this case, Eq. (28) is equal to

$$2C_1D^2(1-\rho)^2 + \frac{\sigma^2\rho^2}{k} \geq \frac{1}{\frac{1}{2C_1D^2} + \frac{k}{\sigma^2}} = O(1).$$

If $2C_1D^2 \geq \frac{\sigma^2\rho}{k(2-\rho)}$, the above is the smallest when $t = T$. In this case, Eq. (28) is at least

$$\frac{\sigma^2\rho}{k(2-\rho)} \geq \frac{\sigma^2\rho}{2k} \geq \frac{\sigma^2 C_1 x^{-1}}{2k}\cdot\frac{1}{\sqrt{T}}. \tag{29}$$

To conclude the proof, we set $x$ so that the right-hand side of (27) and (29) match each other. This leads to

$$x = \sqrt{\frac{2\sigma^2 C_1}{kD^2}}.$$

Thus, by combining the conclusions from both (27) and (29) with this value of $x$, we finally conclude that if $\sum_{i=0}^{T-1}\eta_i \leq \sqrt{T}x^{-1}$, or for all $i = 0, \ldots, T-1$, $\eta_i = \eta < C_1^{-1}$, then in both cases, there exists a function $f$ such that Eq. (26) holds. This completes the proof of Theorem 10. ■

### B.2. A Nonconvex Example

Next, we construct a nonconvex function. Let $e_t$ be the basis vector for the $t$-th dimension, for $t = 0, 1, \ldots, T-1$. We consider a chain-like function:

$$f(W) = \frac{1}{2G}\langle W, e_0\rangle^2 + \sum_{i=0}^{T-1} h_i(\langle W, e_{i+1}\rangle), \tag{30}$$

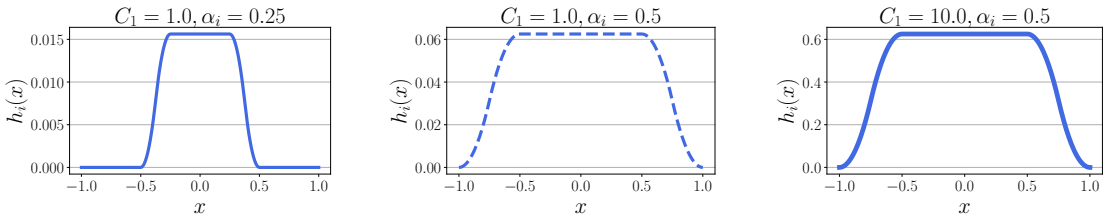


Figure 3: Examples of piecewise quadratic functions $h_i$ with Lipschitzness $C_1$ and parameter $\alpha_i$.

where $h_i$ a quadratic function parametrized by $\alpha_i$, defined as follow:

$$h_i(x) = \begin{cases} \frac{1}{4}C_1\alpha_i^2 & |x| \le \alpha_i, \\ -\frac{1}{2}C_1\big(|x| - \alpha_i\big)^2 + \frac{1}{4}C_1\alpha_i^2 & \alpha_i \le |x| \le \frac{3}{2}\alpha_i, \\ \frac{1}{2}C_1\big(|x| - 2\alpha_i\big)^2 & \frac{3}{2}\alpha_i \le |x| \le 2\alpha_i, \\ 0 & 2\alpha_i \le |x|. \end{cases}$$

See Figure 3 for several examples of this function.

One can verify that $\nabla h_i$ is $C_1$-Lipschitz but is not differentiable. Additionally, $\nabla f$ is $C_1$-Lipschitz when $G \le C_1^{-1}$ (cf. Eq. (30)). As for the other parts of the construction:

- Let $W_0$ satisfy that $\langle W_0, e_i\rangle = 0$ for any $i \ge 1$ while $\langle W_0, e_0\rangle \ne 0$.
- Let $\mathcal{P}$ satisfy that the $i$-th coordinate is at most $2^{-1}\alpha_{i-1}$, for $i = 1, \dots, T$.
- Let the gradient oracle be chosen such that the $i$-th step's gradient noise $\xi_i = \langle \xi_i, e_{i+1}\rangle e_{i+1}$. This implies only coordinate $i+1$ is updated in step $i$, as long as $\langle \xi_i, e_{i+1}\rangle \le 2^{-1}\alpha_i$.

Based on this construction, in step $i$, the gradient noise $\xi_i$ plus the perturbation noise is less than $\alpha_i$ at coordinate $i+1$. Thus, we have that $h_i'(\langle W_t, e_{i+1}\rangle) = 0$. This implies $\nabla f(W_i)$ is equal to $G^{-1}\langle W_i, e_0\rangle$. Roughly speaking, the above construction will go through provided the learning rate $\eta_i$ is smaller than $\alpha_i$. Recall from Assumption (3) that $F(W_0) \le D^2$, which is the same as $\sum_{i=0}^{T-1} \alpha_t^2 \le D^2$. With some calculation, the learning rates must satisfy $\sum_{i=0}^{T-1} \eta_i \lesssim \sqrt{T}$. This ensures the iterates will be trapped close to $W_0$ even after $T$ steps; Thus, the gradient of every iterate must be "large". The details are stated in Lemma 15, App. D.4.

When the learning rate is fixed at $\eta \gtrsim T^{-1/2}$, we construct a chain-like function as above but use a different piece-wise quadratic function with a fixed $\alpha$ (cf. Lemma 16, App. D.4). The gradient noise grows by $1 - \eta/C_1$ up to $T$ steps at each step. We then carefully set $\alpha$ to lower bound the norm of the gradient. Based on these two results, we state the lower bound for a nonconvex example.

**Theorem 11** *Suppose the learning rates $\eta_0, \dots, \eta_{T-1}$ are at most $C_1^{-1}$ and satisfy $\sum_{i=0}^{T-1} \eta_i \lesssim \sqrt{kT}$ or $\eta_i = \eta < C_1^{-1}$ for any epoch $i$. There exists a nonconvex, differentiable, but not twice-differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ with dimension $d \ge T$ satisfying Assumption 3, a gradient oracle satisfying Assumption 1, and perturbation distribution $\mathcal{P}$ such that the following is true*

$$\min_{1 \le t \le T} \mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \ge D\sqrt{\frac{C_1\sigma^2}{32k \cdot T}}. \tag{31}$$

One can see that the rate in Eq. (31) again matches our upper bound for the case of differentiable functions in Eq. (10), when $H(\mathcal{P})$ is a constant (e.g. when $\mathcal{P} = N(0, d^{-1}\,\mathrm{Id}_d)$.

### B.3. Extensions

The lower bound regarding convex functions can be naturally extended to momentum SGD:

$$V_{t+1} = \mu V_t - \eta_t G_t, \text{ and } M_{t+1} = M_t + V_{t+1}, \tag{32}$$

where $G_t$ refers to the stochastic gradient in Eq. (2). This can be seen by generalizing the argument of Lemma 8 to weighted combinations of iterates. We state the following corollary, leaving its proof to Appendix D.5.

**Corollary 12** *There exists a convex quadratic function satisfying Assumption 3 with Lipschitz Hessians such that for $M_1, \ldots, M_T$ following momentum SGD, the following must be true:*

$$\min_{1 \le t \le T} \mathbb{E}\left[\|\nabla F(M_t)\|^2\right] \gtrsim D\sqrt{\frac{C_1 \sigma^2}{k \cdot T}}. \tag{33}$$

Next, we discuss extensions to SGD with adaptive learning rates. Again, notice that the proof of Lemma 9 holds for any arbitrary learning rates. Furthermore, the proof of Theorem 11 also applies to any learning rates under certain bounded conditions. It is conceivable that similar results can be derived for SGD with adaptive step sizes under similar bounded conditions when $\eta_i$ is a function of previous weights and step size.

The construction for the nonconvex case can also be extended to SGD with adaptive learning rates such as AdaGrad. To see this, notice that the proof of Lemma 15 does not use any specific properties of the learning rates. In fact, we can set the width of each $h_t$ function, $\alpha_t$, in proportion to the learning rate $\eta_t$, for an arbitrary $t$. Thus, even if the algorithm uses an adaptive learning rate schedule such as AdaGrad, the construction can still go through.

## Appendix C. Experiment Details

We describe setups and results left from the main text. We consider two types of image classification datasets. First, we use two medical image datasets related to diabetic retinopathy classification using eye fundus images. Each dataset contains the interior surface images of a single eye and labels regarding the severity of diabetic retinopathy of the eye. The two datasets are available in the following sources: Retina I (Messidor) [1] and Retina II (APTOS) [2]. Second, we consider fine-grained image classification datasets, including Aircrafts [3], Birds[4], and Indoor[5]. Besides, we use the following datasets to illustrate the consistent performance of our algorithm, including Flowers, Cars, Caltech-256, CIFAR-10, and CIFAR-100. We report the dataset statistics in Table 2.

**Results on CIFAR.** We report fine-tuning results of ResNet-34 on CAFAR-10 and CIFAR-100. Our algorithm achieves 18% and 26% smaller the Hessian eigenvalues and traces, respectively. The results are shown in Table 4.

1. https://www.adcis.net/en/third-party/messidor2/
2. https://kaggle.com/competitions/aptos2019-blindness-detection
3. https://www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/
4. http://www.vision.caltech.edu/visipedia/CUB-200-2011.html
5. http://web.mit.edu/torralba/www/indoor.html

Table 2: Basic statistics for image classification tasks.

| | Retina I | Retina II | Aircrafts | Birds | Indoor | Flowers | Cars | Caltech-256 | CIFAR-10 | CIFAR-100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Training | 1,396 | 2,930 | 3,334 | 5,395 | 4,824 | 1,020 | 7,330 | 7,680 | 45,000 | 45,000 |
| Validation | 248 | 732 | 3,333 | 599 | 536 | 1,020 | 814 | 5,120 | 5,000 | 5,000 |
| Test | 250 | 733 | 3,333 | 574 | 1,340 | 6,149 | 8,441 | 5,120 | 10,000 | 10,000 |
| Classes | 5 | 5 | 100 | 200 | 67 | 102 | 196 | 256 | 10 | 100 |

Table 3: Top panel: Largest eigenvalue of neural net loss Hessian matrix. Middle panel: Hessian trace. Lower panel: Test accuracy. All results are averaged over three random seeds.

| Dataset | Retina I | Retina II | Aircrafts | Birds | Indoor |
|---|---|---|---|---|---|
| SGD | 8,996±92 | 2,413±32 | 1,239±43 | 1,286±76 | 1,169±66 |
| LS | 4,911±73 | 1,494±99 | 1,339±84 | 992±55 | 876±81 |
| SAM | 4,246±55 | 1,341±26 | 958±60 | 1,061±49 | 975±56 |
| ASAM | 4,267±58 | 1,629±41 | 641±89 | 1,089±56 | 722±56 |
| NSO (Alg. 1) | **3,916**±47 | **1,095**±49 | **612**±44 | **1,045**±68 | **688**±58 |
| SGD | 30,433±217 | 6,921±54 | 6,218±63 | 4,252±93 | 4,078±78 |
| LS | 19,219±119 | 4,559±49 | 6,332±76 | 4,585±79 | 4,196±36 |
| SAM | 16,411±161 | 3,702±94 | 5,034±59 | 3,515±48 | 3,789±49 |
| ASAM | 14,745±131 | 4,174±52 | 5,191±32 | 3,799±44 | 3,124±73 |
| NSO (Alg. 1) | **11,554**±77 | **3,519**±57 | **4,193**±46 | **3,067**±65 | **2,991**±32 |
| SGD | 61.69±0.89 | 83.70±0.53 | 59.76±0.71 | 73.34±0.11 | 75.95±0.46 |
| LS | 63.56±0.72 | 83.65±0.68 | 58.47±0.29 | 74.38±0.33 | 75.87±0.35 |
| SAM | 64.39±0.66 | 84.65±0.78 | 61.54±0.82 | 73.65±0.65 | 76.57±0.56 |
| ASAM | 64.79±0.37 | 84.88±0.38 | 62.02±0.61 | 74.72±0.58 | 76.72±0.31 |
| NSO (Alg. 1) | **66.57**±0.75 | **85.24**±0.45 | **62.31**±0.36 | **75.35**±0.53 | **77.39**±0.55 |

Table 4: Comparisons on CIFAR datasets. **Left**: Largest eigenvalue of the loss Hessian matrix. **Middle**: Hessian trace. **Right**: Test accuracy. All results are averaged over 3 random seeds.

| | Largest Eigenvalue | | Trace | | Test Accuracy | |
|---|---|---|---|---|---|---|
| | CIFAR-10 | CIFAR-100 | CIFAR-10 | CIFAR-100 | CIFAR-10 | CIFAR-100 |
| Vanilla SGD | 1,522 ± 45 | 4,870 ± 88 | 4,738 ± 73 | 14,372 ± 346 | 95.48±0.10 | 82.25±0.16 |
| Label smoothing | 1,416 ± 45 | 3,459 ± 95 | 2,903 ± 92 | 11,295 ± 384 | 96.68±0.10 | 83.75±0.19 |
| SAM | 1,372 ± 49 | 3,419 ± 75 | 2,847 ± 78 | 10,230 ± 385 | 96.64±0.47 | 83.48±0.16 |
| ASAM | 1,425 ± 67 | 2,599 ± 79 | 2,830 ± 57 | 10,478 ± 263 | 96.73±0.16 | 83.84±0.12 |
| NSO (Alg. 1) | **1,136 ± 58** | **2,184 ± 59** | **2,061 ± 70** | **5,829 ± 211** | **97.10±0.28** | **84.29±0.27** |

## Appendix D. Proofs for Technical Lemmas

### D.1. Proof of Lemma 6

The smoothness condition in Assumption 3 is equivalent to the following domination inequality:

$$|f(W_2) - f(W_1) - \langle \nabla f(W_1), W_2 - W_1 \rangle| \leq \frac{C_1}{2} \|W_2 - W_1\|^2 . \tag{34}$$

See, e.g., Bach [2, Chapter 5]. We state a similar result on the Hessian as follows.

**Proposition 13** *Suppose $f(W)$ is twice-differentiable and the Hessian of $f(W)$ is $C_2$-Lipschitz. Then, for any $W_1 \in \mathbb{R}^d$ and $W_2 \in \mathbb{R}^d$, the following holds:*

$$\left\|\nabla f(W_2) - \nabla f(W_1) - \mathbf{H}\big[f(W_1)\big](W_2 - W_1)\right\| \leq \frac{C_2}{2}\left\|W_2 - W_1\right\|^2. \tag{35}$$

**Proof** We apply a line integral along $W_2 - W_1$ below:

$$\begin{aligned}
&\left\|\nabla f(W_2) - \nabla f(W_1) - \mathbf{H}\big[f(W_1)\big](W_2 - W_1)\right\| \\
=&\left\|\int_0^1 \mathbf{H}\big[f(W_1 + \tau(W_2 - W_1))\big]d\tau(W_2 - W_1) - \mathbf{H}\big[f(W_1)\big](W_2 - W_1)\right\| \\
=&\left\|\int_0^1 \Big(\mathbf{H}\big[f(W_1 + \tau(W_2 - W_1))\big] - \mathbf{H}\big[f(W_1)\big]\Big)(W_2 - W_1)d\tau\right\| \\
\leq&\int_0^1 C_2\tau\left\|W_2 - W_1\right\|^2 d\tau = \frac{C_2}{2}\left\|W_2 - W_1\right\|^2.
\end{aligned}$$

■

Based on the above result, we provide proofs for the variance of $\delta_i$ and $\xi_i$ for any $i = 0, 1, \ldots, T-1$.

**Proof of Lemma 6** First, we can see that

$$\begin{aligned}
\mathop{\mathbb{E}}_{U_i^1,\ldots,U_i^k}\left[\|\delta_i\|^2\right] &= \mathop{\mathbb{E}}_{U_i^1,\ldots,U_i^k}\left[\left\|\frac{1}{2k}\sum_{j=1}^k\Big(\nabla f(W_i + U_i^j) + \nabla f(W_i - U_i^j) - 2\nabla F(W_i)\Big)\right\|^2\right] \\
&= \frac{1}{k^2}\sum_{j=1}^k \mathop{\mathbb{E}}_{U_i^j}\left[\left\|\frac{1}{2}\Big(\nabla f(W_i + U_i^j) + \nabla f(W_i - U_i^j) - 2\nabla F(W_i)\Big)\right\|^2\right] \\
&\qquad\qquad (U_i^{j_1} \perp U_i^{j_2} \text{ when } j_1 \neq j_2) \\
&= \frac{1}{k}\mathop{\mathbb{E}}_{U_i^1}\left[\left\|\frac{1}{2}\Big(\nabla f(W_i + U_i^1) + \nabla f(W_i - U_i^1)\Big) - \nabla F(W_i)\right\|^2\right] \\
&\qquad\qquad (U_i^1, \ldots, U_i^k \text{ are identical})
\end{aligned}$$

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

$$\Pr[U \in A, V \in B] = \Pr[U \in A] \cdot \Pr[V \in B], \forall A, B \in \text{Borel}(\mathbb{R}).$$

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

$$\mathbb{E}\left[\|\delta_i\|^2\right] = \frac{1}{k}\mathop{\mathbb{E}}_{U\sim\mathcal{P}}\left[\left\|\frac{1}{2}\Big(\nabla f(W_i+U)+f(W_i-U)\Big)-\nabla F(W_i)\right\|^2\right]. \tag{36}$$

Next, we deal with the variance of the two-point stochastic gradient, depending on whether $f$ has Lipschitz-Hessian or not. If $f$ has $C_1$-Lipschitz gradient, we will show that

$$\mathop{\mathbb{E}}_{U}\left[\left\|\frac{1}{2}\Big(\nabla f(W+U)+\nabla f(W-U)\Big)-\nabla F(W)\right\|^2\right] \le 2C_1^2 H(\mathcal{P}). \tag{37}$$

By equation (34) from Proposition 13, the left-hand side of equation (37) is equal to

$$\begin{aligned}
&\mathop{\mathbb{E}}_{U}\left[\left\|\frac{1}{2}\Big(\nabla f(W+U)-\nabla F(W)\Big)+\frac{1}{2}\Big(\nabla f(W-U)-\nabla F(W)\Big)\right\|^2\right]\\
\le&\mathop{\mathbb{E}}_{U}\left[\frac{1}{2}\|\nabla f(W+U)-\nabla F(W)\|^2+\frac{1}{2}\|\nabla f(W-U)-\nabla F(W)\|^2\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(by Cauchy-Schwartz inequality)}\\
=&\frac{1}{2}\mathop{\mathbb{E}}_{U}\left[\|\nabla f(W+U)-\nabla F(W)\|^2\right] \qquad\text{(by symmetry of } \mathcal{P} \text{ since it has mean zero)}\\
=&\frac{1}{2}\mathop{\mathbb{E}}_{U}\left[\left\|\mathop{\mathbb{E}}_{U'\sim\mathcal{P}}\left[\nabla f(W+U)-\nabla f(W+U')\right]\right\|^2\right]\\
\le&\frac{1}{2}\mathop{\mathbb{E}}_{U}\left[\mathop{\mathbb{E}}_{U'\sim\mathcal{P}}\left[\|\nabla f(W+U)-\nabla f(W+U')\|^2\right]\right]\\
\le&\frac{1}{2}\mathop{\mathbb{E}}_{U,U'}\left[C_1^2\|U-U'\|^2\right]=\frac{1}{2}C_1^2\mathop{\mathbb{E}}_{U,U'}\left[\|U\|^2+\|U'\|^2\right]=C_1^2H(\mathcal{P}) \qquad\text{(by equation (34))}
\end{aligned}$$

If, in addition, the Hessian of $f$ is $C_2$-Lipschitz, we show that

$$\mathop{\mathbb{E}}_{U}\left[\left\|\frac{1}{2}\Big(\nabla f(W+U)+\nabla f(W-U)\Big)-\nabla F(W)\right\|^2\right] \le \frac{3}{8}C_2^2 G(\mathcal{P}). \tag{38}$$

By equation (35) from Proposition 13, we know that

$$\left\|\nabla f(W+U)-\nabla f(W)-\mathbf{H}\big[f(W)\big]U\right\| \le \frac{C_2}{2}\|U\|^2, \text{ and} \tag{39}$$

$$\left\|\nabla f(W-U)-\nabla f(W)+\mathbf{H}\big[f(W)\big]U\right\| \le \frac{C_2}{2}\|U\|^2. \tag{40}$$

Next,

$$\begin{aligned}
\mathop{\mathbb{E}}_{U}\left[\|\nabla F(W)-\nabla f(W)\|^2\right] &= \|\nabla F(W)-\nabla f(W)\|^2 && (41)\\
&= \left\|\mathop{\mathbb{E}}_{U}\left[\nabla f(W+U)-\nabla f(W)-\mathbf{H}\big[f(W)\big]U\right]\right\|^2 \\
&\qquad\qquad \text{(since } \mathbb{E}_U\left[U\right]=0\text{)} \\
&\le \mathop{\mathbb{E}}_{U}\left[\left\|\nabla f(W+U)-\nabla f(W)-\mathbf{H}\big[f(W)\big]U\right\|^2\right] \\
&\qquad\qquad \text{(by Cauchy-Schwartz ineq.)} \\
&\le \mathop{\mathbb{E}}_{U}\left[\left(\frac{C_2}{2}\|U\|^2\right)^2\right], && (42)
\end{aligned}$$

where the last step uses equation (40). Hence, using triangle inequality, the left-hand side of equation (38) is equal to

$$\begin{aligned}
&\frac{1}{k}\mathop{\mathbb{E}}_{U}\left[\left\|\frac{1}{2}\nabla f(W+U)-\frac{1}{2}\nabla F(W)+\frac{1}{2}\nabla f(W-U)-\frac{1}{2}\nabla f(W)+\nabla f(W)-\nabla F(W)\right\|^2\right] \\
\le&\frac{1}{k}\left(\mathop{\mathbb{E}}_{U}\left[\left(\frac{C_2}{4}\|U\|^2\right)^2+\left(\frac{C_2}{4}\|U\|^2\right)^2+\left(\frac{C_2}{2}\|U\|^2\right)^2\right]\right) \\
&\qquad\qquad \text{(by combining equations (39), (40), and (42))} \\
=&\frac{3C_2^2}{8k}\cdot\mathop{\mathbb{E}}_{U}\left[\|U\|^4\right]=\frac{3C_2^2}{8k}G(\mathcal{P}).
\end{aligned}$$

Therefore, we have proved equation (38).

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

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

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 $g_z(\cdot)$ from equation (3), Thus, the proof is finished. ■

### D.2. Proof of Lemma 7

In this result, we analyze the convergence of NSO for functions with Lipschitz gradients.

**Proof of Lemma 7** Using Proposition 4 and Proposition 13, we have

$$
\begin{aligned}
&F(W_{i+1}) \\
\leq& F(W_i) + \langle \nabla F(W_i), W_{i+1} - W_i \rangle + \frac{C_1}{2}\eta_i^2 \left\| \frac{1}{2}\Big(\nabla f(W_i + U_i) + \nabla f(W_i - U_i)\Big) + \xi_i \right\|^2 \\
=& F(W_i) - \eta_i \langle \nabla F(W_i), \delta_i + \xi_i + \nabla F(W_i) \rangle + \frac{C_1 \eta_i^2}{2} \left\| \delta_i + \xi_i + \nabla F(W_i) \right\|^2 \\
=& F(W_i) - \Big(\eta_i - \frac{C_1\eta_i^2}{2}\Big) \left\| \nabla F(W_i) \right\|^2 - \Big(\eta_i - C_1\eta_i^2\Big) \langle \nabla F(W_i), \delta_i + \xi_i \rangle + \frac{C_1\eta_i^2}{2} \left\| \delta_i + \xi_i \right\|^2 .
\end{aligned}
$$

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

$$
\begin{aligned}
&\sum_{i=0}^{T-1} \Big(\eta_i - \frac{C_1\eta_i^2}{2}\Big) \left\| \nabla F(W_i) \right\|^2 \qquad (43) \\
\leq& F(W_0) - F(W_T) - \sum_{i=0}^{T-1} \Big(\eta_i - C_1\eta_i^2\Big) \langle \nabla F(W_i), \delta_i + \xi_i \rangle + \frac{C_1}{2} \sum_{i=0}^{T-1} \eta_i^2 \left\| \delta_i + \xi_i \right\|^2 \\
\leq& D^2 - \sum_{i=0}^{T-1} \Big(\eta_i - C_1\eta_i^2\Big) \langle \nabla F(W_i), \delta_i + \xi_i \rangle + \frac{C_1}{2} \sum_{i=0}^{T-1} \eta_i^2 \left\| \delta_i + \xi_i \right\|^2 . \qquad (44)
\end{aligned}
$$

where $F(W_0) - F(W_T) \leq F(W_0) - \min_{W \in \mathbb{R}^d} F(W) \leq D^2$. For any $t = 0, 1, \ldots, T-1$, clearly, $\eta_t \leq 2\eta_t - C_1\eta_t^2$ as long as $0 < \eta_t \leq \frac{1}{C_1}$. Hence, we have

$$
\frac{1}{2} \sum_{t=0}^{T-1} \eta_t \left\| \nabla F(W_t) \right\|^2 \leq \sum_{t=0}^{T-1} \Big(\eta_t - \frac{C_1\eta_t^2}{2}\Big) \left\| \nabla F(W_t) \right\|^2 ,
$$

which implies that

$$
\frac{1}{2} \sum_{i=0}^{T-1} \eta_i \left\| \nabla F(W_i) \right\|^2 \leq D^2 - \sum_{i=0}^{T-1} \Big(\eta_i - C_1\eta_i^2\Big) \langle \nabla F(W_i), \delta_i + \xi_i \rangle + \frac{C_1}{2} \sum_{i=0}^{T-1} \eta_i^2 \left\| \delta_i + \xi_i \right\|^2 . \qquad (45)
$$

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

$$
\mathop{\mathbb{E}}_{U_t} [\delta_t] = \frac{1}{2} \mathop{\mathbb{E}}_{U_t} [\nabla f(W_t - U_t) - \nabla f(W_t + U_t)] = 0. \qquad (46)
$$

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 7. We can write equation (45) equivalently as

$$\begin{aligned}\mathop{\mathbb{E}}_{t;\, U_0,\ldots,U_{T-1},\xi_0,\xi_1,\ldots,\xi_{T-1}}\left[\|\nabla F(W_t)\|^2\right] &= \frac{\sum_{i=0}^{T-1}\eta_i\,\mathbb{E}\left[\|\nabla F(W_i)\|^2\right]}{\sum_{i=0}^{T-1}\eta_i}\\ &\le \frac{2D^2 + C_1\sum_{i=0}^{T-1}\eta_i^2\,\mathbb{E}\left[\|\delta_i+\xi_i\|^2\right]}{\sum_{i=0}^{T-1}\eta_i}\\ &= \frac{2D^2 + C_1\sum_{i=0}^{T-1}\eta_i^2\left(\mathbb{E}\left[\|\delta_i\|^2\right]+\mathbb{E}\left[\|\xi_i\|^2\right]\right)}{\sum_{i=0}^{T-1}\eta_i}.\end{aligned}$$

where we use the fact that $\delta_i$ and $\xi_i$ are independent for any $i$. Hence, we have finished the proof of equation (16). ■

**Remark 14** *Our analysis builds on the work of Ghadimi and Lan [14]. The difference is that we additionally deal with stochasticity due to random perturbations. Additionally, we analyze the case of injecting multiple perturbations, which requires a careful analysis of the variance of NSO.*

### D.3. Proof of Lemma 9

We know that $F(W) = f(W) + d$ if $\mathcal{P}$ is the isotropic Gaussian distribution. Clearly, the gradient of $F$ is $C_1$-Lipschitz, and the Hessian of $F$ is arbitrarily Lipschitz continuous. Let the initialization $W_0 \in \mathbb{R}^d$ be set such that $F(W_0) - \min_{W\in\mathbb{R}^d} F(W) = D^2$. This condition can be met we can set $W_0$ as a vector whose Euclidean norm is equal to $D\sqrt{2\max\left\{C_1^{-1}, 2\sum_{i=0}^{T-1}\eta_i\right\}}$.

**Proof of Lemma 9** Clearly, the norm of the gradient of $F(W)$ is equal to

$$\|\nabla F(W)\| = \frac{1}{\kappa}\|W\|. \tag{47}$$

Following the update rule in SGD (cf. equation (2)), similar to equation (19), $W_t$ evolves as follows:

$$W_{t+1} = \left(1 - \frac{\eta_t}{\kappa}\right)W_t - \eta_t\bar{\xi}_t, \tag{48}$$

where $\bar{\xi}_t$ has variance equal to $\sigma^2/k$, according to the proof of Lemma 8. By iterating equation (48) from the initialization, we can get a closed-form equation for $W_t^{(1)}$, for any $t = 1, 2, \ldots, T$:

$$W_t = W_0\prod_{j=0}^{t-1}\left(1-\frac{\eta_j}{\kappa}\right) - \sum_{k=0}^{t-1}\eta_k\xi_k\prod_{j=k+1}^{t-1}\left(1-\frac{\eta_j}{\kappa}\right). \tag{49}$$

Following equation (47), we can show that

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

Thus, in expectation,

$$\begin{aligned}
\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] &= \kappa^{-2}\,\mathbb{E}\left[\|W_t\|^2\right] \\
&= \kappa^{-2}\,\|W_0\|^2 \prod_{j=0}^{t-1}\left(1-\kappa^{-1}\eta_j\right)^2 + \kappa^{-2}\sum_{i=0}^{t-1}\mathbb{E}\left[\left(\eta_i\bar{\xi}_i\prod_{j=i+1}^{t-1}\left(1-\kappa^{-1}\eta_j\right)\right)^2\right] \\
&= \kappa^{-2}\,\|W_0\|^2 \prod_{j=0}^{t-1}\left(1-\kappa^{-1}\eta_j\right)^2 + \kappa^{-2}\sum_{i=0}^{t-1}\eta_i^2\prod_{j=i+1}^{t-1}\left(1-\kappa^{-1}\eta_j\right)^2\mathbb{E}\left[\|\bar{\xi}_i\|^2\right] \\
&= 2D^2\kappa^{-1}\prod_{j=0}^{t-1}\left(1-\kappa^{-1}\eta_j\right)^2 + \frac{\sigma^2\kappa^{-2}}{k}\sum_{i=0}^{t-1}\eta_i^2\prod_{j=i+1}^{t-1}\left(1-\kappa^{-1}\eta_j\right)^2, \qquad (50)
\end{aligned}$$

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 (50), we note that for all $z \in [0,1]$,

$$1-\frac{z}{2} \geq \exp\left(\log\frac{1}{2}\cdot z\right). \qquad (51)$$

Hence, applying equation (51) to the right-hand side of equation (50), we obtain that for any $i = 0,1,\ldots,t-1$,

$$\begin{aligned}
&\prod_{j=i}^{t-1}\left(1-\frac{\eta_j}{\max\{C_1^{-1}, 2\sum_{j=i}^{T-1}\eta_i\}}\right) \\
&\geq \exp\left(\log\frac{1}{2}\cdot\sum_{j=i}^{t-1}\frac{\eta_j}{\max\{(2C_1)^{-1}, \sum_{i=0}^{T-1}\eta_i\}}\right) \geq \frac{1}{2}.
\end{aligned}$$

Thus, equation (50) must be at least

$$\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \geq 2D^2\kappa^{-1}\cdot\frac{1}{4} + \frac{\sigma^2\kappa^{-2}}{k}\sum_{i=0}^{t-1}\eta_i^2\cdot\frac{1}{4}. \qquad (52)$$

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

$$\min_{1\leq t\leq T}\mathbb{E}\left[\|\nabla F(W_t)\|^2\right] \geq \frac{D^2}{2\max\{C_1^{-1}, 2\sum_{i=0}^{T-1}\eta_i\}}.$$

Thus, the proof of Lemma 9 is finished. ■

### D.4. Supporting Lemmas for Theorem 11

For technical reasons, 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 $\tilde{U}$.

**Lemma 15** *In the setting of Theorem 11, suppose the learning rates satisfy that* $\sum_{i=0}^{T-1}\eta_i \leq \sqrt{\frac{D^2kT}{2\sigma^2C_1}}$, *consider the function* $f(W)$ *constructed in equation* (30)*, we have that the conclusion of equation* (31) *holds.*

**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

$$\xi_t = \langle \xi_t, e_{t+1} \rangle \text{ and } |\langle \xi_t, e_{t+1} \rangle| \leq \frac{\sigma}{\sqrt{k}}, \tag{53}$$

where $e_{t+1}$ is a basis vector whose $(t+1)$-th entry is one and otherwise is zero. In words, the above notation means that only the $(t+1)$-th coordinate of $\xi_t$ is nonzero. 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 $\xi_t^{(i)}$ also satisfies the above condition (53). Thus, we have

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

We consider the objective function $f(W) : \mathbb{R}^d \to \mathbb{R}$ defined in equation (30), Section B. Here we require $d \geq T+1$ and $\alpha_i = 2\eta_i \sigma / \sqrt{k}$. The proof is very similar to Lemma 9. Let $G = \max\left\{C_1^{-1}, 2\sum_{i=0}^{T-1} \eta_i\right\}$. We analyze the dynamics of two-point SGD (Algorithm 1) with the objective function $f(W)$ and the starting point $W_0 = D\sqrt{G} \cdot e_0$. For the first iteration, we have

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

where $U$ is a truncated distribution of $\mathcal{P} \sim N(0, \mathrm{Id}_d)$ with $\langle U, e_i \rangle = \min\{\mathcal{P}_i, a_i\}$ and $a_i = \eta_{i-1}\sigma/\sqrt{k}$. Next, from the properties of $h_1$, we get

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

Here, using the fact that $\alpha_0 = 2\eta_0\sigma/\sqrt{k}$, $|\langle U, e_1 \rangle| \leq \eta_0 \sigma/\sqrt{k}$, and $\langle \bar{\xi}_0, e_1 \rangle \leq \sigma/\sqrt{k}$, we obtain

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

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 $\alpha_i = 2\eta_i\sigma/\sqrt{k}$ and the fact that $|\langle U, e_{i+1} \rangle| \leq \eta_i \sigma/\sqrt{k}$, which renders the gradient as zero, for any $i = 1, 2, \ldots, T-1$. At the $t$-th iteration, suppose 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 have

$$\begin{aligned} W_{t+1} &= W_t - \eta_t\Big(G^{-1}\langle W_t, e_0\rangle + \bar{\xi}_t\Big) \\ &= W_0 \prod_{i=0}^{t}\Big(1-\eta_i G^{-1}\Big) - \sum_{i=0}^{t}\eta_i\bar{\xi}_i. \end{aligned} \tag{54}$$

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

$$\begin{aligned} F(W_0) - \min_{W\in\mathbb{R}^d} F(W) &= F(W_0) \\ &= \frac{1}{2G}(D\sqrt{G})^2 + \sum_{i=0}^{T-1}\frac{C_1}{4}\big(2\eta_i\sigma/\sqrt{k}\big)^2 \le D^2, \\ &\qquad\qquad \text{(since } \langle W_0+U, e_{i+1}\rangle \le \alpha_i) \end{aligned}$$

which implies that we should set the learning rates to satisfy

$$\frac{1}{T}\Big(\sum_{i=0}^{T-1}\eta_i\Big)^2 \le \sum_{i=0}^{T-1}\eta_i^2 \le \frac{kD^2}{2C_1\sigma^2}. \tag{55}$$

Applying equation (51) to the right-hand side of equation (54), we obtain that for any $t$,

$$\prod_{i=0}^{t}\Big(1-\eta_i G^{-1}\Big) \ge \frac{1}{2}.$$

From equation (54), we conclude that

$$\begin{aligned} \min_{1\le i\le T}\|\nabla F(W_i)\|^2 = \min_{1\le i\le T}\Big(G^{-1}\langle W_t, e_0\rangle\Big)^2 &\ge \frac{1}{4}G^{-2}(D\sqrt{G})^2 \\ &= \frac{D^2}{4}\min\Big\{C_1, \frac{1}{2\sum_{i=0}^{T-1}\eta_i}\Big\} \\ &\ge \frac{D^2}{4}\min\Big\{C_1, \frac{\sqrt{2C_1\sigma^2}}{2D\sqrt{kT}}\Big\} \\ &\ge D\sqrt{\frac{C_1\sigma^2}{32kT}}. \end{aligned}$$

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 (31) holds for $W_i$ for any $i = 1, 2, \ldots, T$. The proof is finished. ■

Next, we consider the case of large, fixed learning rates.

**Lemma 16** *In the setting of Theorem 11, suppose the learning rates satisfy that* $\sum_{i=0}^{T-1}\eta_i \ge \sqrt{\frac{D^2kT}{2\sigma^2C_1}}$ *and* $\eta_i = \eta$ *for some fixed* $\eta \le C_1^{-1}$*. Then, consider the function from equation (30), we have that equation (31) holds.*

**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:

$$g(x) = \begin{cases} -\frac{C_1}{2}x^2 + \frac{C_1}{4}\alpha^2 & |x| \le \alpha/2, \\ \frac{C_1}{2}(|x| - \alpha)^2 & \alpha/2 \le |x| \le \alpha, \\ 0 & \alpha \le |x|. \end{cases}$$

One can verify that $g$ has $C_1$-Lipschitz gradient, but $g$ is not twice-differentiable. We also consider a chain-like function:

$$f(W) = g(\langle W, e_0 \rangle) + \sum_{t=0}^{d-1} \frac{C_1}{2} \langle W, e_{t+1} \rangle^2 \tag{56}$$

From the definition of $f$, $f$ also has $C_1$-Lipschitz gradient. Similar to (53), 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

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

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

$$|\langle \bar{\xi}_t, e_{t+1} \rangle| \le c\sigma \text{ and } \mathbb{E}\left[\left\|\bar{\xi}_t\right\|^2\right] \le \frac{\sigma^2}{k} \tag{57}$$

Next, we analyze the dynamics of two-point SGD (Algorithm 1) with the objective function $f(W)$ and the starting point $W_0 = \sum_{i=1}^d \sqrt{\frac{D^2}{C_1 d}} \cdot e_i$. In this case, by setting $\eta_i = \eta$ for all $i = 0, 1, \ldots, T-1$. Recall that $\eta < C_1^{-1}$. Denote by $\rho = C_1\eta$, which is strictly less than one.

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

$$\begin{aligned} W_1 &= W_0 - \eta\Big(\frac{1}{2}\big(\nabla f(W_0 + U) + \nabla f(W_0 - U)\big) + \bar{\xi}_0\Big) \\ &= (1 - C_1\eta)W_0 - \eta\bar{\xi}_0. \end{aligned}$$

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

$$g'(\eta\langle \bar{\xi}_0, e_0 \rangle + \langle U, e_0 \rangle) + g'(\eta\langle \bar{\xi}_0, e_0 \rangle - \langle U, e_0 \rangle) = -2C_1\eta\langle \bar{\xi}_0, e_0 \rangle.$$

Then, in the next iteration,

$$\begin{aligned} W_2 &= W_1 - \eta\Big(C_1 \sum_{i=1}^{d} \langle W_1, e_i \rangle - C_1\eta\bar{\xi}_0 + \bar{\xi}_1\Big) \\ &= (1 - C_1\eta)^2 W_0 - (1 - C_1\eta)\eta\bar{\xi}_0 - \eta\bar{\xi}_1. \end{aligned}$$

Similarly, we use the fact that $\alpha = \frac{1-\rho^T}{1-\rho} \cdot 2c\eta\sigma$ and the fact that $|\langle U, e_0\rangle| \leq c\eta\sigma$, which renders the gradient as $g'(x) = -C_1 x$, for any $i = 1, 2, \ldots, T-1$. At the $t$-th iteration, suppose that

$$W_t = (1 - C_1\eta)^t W_0 - \sum_{i=0}^{t-1} (1 - C_1\eta)^{t-1-i} \eta \bar{\xi}_i.$$

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

$$\begin{aligned} W_{t+1} &= W_t - \eta\Big(C_1 \sum_{i=1}^{d} \langle W_t, e_i\rangle - C_1 \sum_{i=0}^{t-1} (1 - C_1\eta)^{t-1-i} \eta \bar{\xi}_i + \bar{\xi}_t\Big) \\ &= (1 - C_1\eta)^{t+1} W_0 - \sum_{i=0}^{t} (1 - C_1\eta)^{t-1-i} \eta \bar{\xi}_i. \end{aligned} \tag{58}$$

Applying equation (51) to the right-hand side of equation (54), we obtain that for any $t$, Next, from the definition of $g$ above, we have that

$$\begin{aligned} F(W_0) - \min_{W \in \mathbb{R}^d} F(W) &= F(W_0) \\ &= \frac{dC_1}{2}\Big(\sqrt{\frac{D^2}{C_1 d}}\Big)^2 + \frac{C_1}{4}\Big(\frac{2(1-\rho^T)c\eta\sigma}{(1-\rho)}\Big)^2 \leq D^2, \\ &\qquad\qquad \text{(since } \langle W_0 + U, e_0\rangle \leq \alpha) \end{aligned}$$

which implies

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

We conclude that

$$\begin{aligned} \min_{1\leq i\leq T} \mathbb{E}\left[\|\nabla F(W_i)\|^2\right] &= \min_{1\leq i\leq T} \mathbb{E}\left[\sum_{j=1}^{d} C_1^2 \langle W_i, e_j\rangle^2 + C_1^2 \langle W_i, e_0\rangle^2\right] \\ &= \min_{1\leq i\leq T} \Big(dC_1^2(1-\rho)^{2t}\Big(\sqrt{\frac{D^2}{C_1 d}}\Big)^2 + \frac{\sigma^2}{k} \cdot \rho^2 \sum_{i=0}^{t} (1-\rho)^{2(t-1-i)}\Big) \\ &\geq \min_{1\leq i\leq T} \Big(C_1 D^2 (1-\rho)^{2t} + \frac{\sigma^2}{k}\frac{\rho}{2-\rho}\big(1 - (1-\rho)^{2t}\big)\Big) \\ &\geq \min\Big\{C_1 D^2, \frac{\sigma^2}{k}\frac{\rho}{2-\rho}\Big\} \\ &\geq \frac{\sigma^2}{k} C_1 \sqrt{\frac{kD^2}{2T\sigma^2 C_1}} \Big/ \Big(2 - C_1\sqrt{\frac{kD^2}{2T\sigma^2 C_1}}\Big) \geq D\sqrt{\frac{C_1\sigma^2}{16k \cdot T}}. \end{aligned}$$

The proof is finished. ■

### D.5. Proof of Corollary 12

**Proof of Corollary 12** The analysis is similar to Theorem 11. First, consider a quadratic function

$$f(W) = \frac{1}{2C_1} \|W\|^2 .$$

Clearly, $f(W)$ is $C_1$-Lipschitz. Further, $F(W) = f(W) + d$, for $\mathcal{P}$ being the isotropic Gaussian. Let $W_0$ be a vector whose Euclidean norm equals $D\sqrt{2C_1}$. Thus, $F(W_0) - \min_{W \in \mathbb{R}^d} F(W) = D^2$.

As for the dynamic of momentum SGD, recall that

$$V_{t+1} = \mu V_t - \eta_t G_t \text{ and } M_{t+1} = M_t + V_{t+1}.$$

Let $\Delta_t = M_t - W_t$, i.e., this is the difference between the iterate of vanilla SGD and momentum SGD, for every $t = 1, 2, \ldots, T$. Thus, we have that

$$\begin{aligned}
\Delta_{t+1} &= M_{t+1} - W_{t+1} \\
&= M_t + V_{t+1} - (W_t - \eta_t G_t) \\
&= \Delta_t + \mu V_t.
\end{aligned}$$

This implies that $\Delta_t = \sum_{i=0}^{t-1} \mu^{t-i} V_i$. Next, we can see that

$$V_{t+1} = \mu V_t - (W_{t+1} - W_t).$$

By applying this through for $t = 0, 1, \ldots$ and so on, we can obtain that

$$V_{t+1} = W_{t+1} - \sum_{i=1}^{t} (\mu^{t-i} - \mu^{t-i+1}) W_i - \mu^t W_0.$$

As a result,

$$\begin{aligned}
\Delta_t = M_t - W_t &= \sum_{j=0}^{t-j} \mu^{t-j} V_j \\
&= \sum_{j=0}^{t-j} \mu^{t-j} \left( W_j - \sum_{i=1}^{j-1} (\mu^{j-1-i} - \mu^{j-i}) W_i - \mu^{j-1} W_0 \right) \\
&= \sum_{j=0}^{t-j} \left( \mu^{t-j} W_j - \mu^{j-1} W_0 - \sum_{i=1}^{j-1} (\mu^{t-i-1} - \mu^{t-i}) W_i \right). \qquad (59)
\end{aligned}$$

Thus, we can see that $M_t$ is now a linear combination of $W_t, W_{t-1}, \ldots, W_0$, and the coefficients of this combination are determined based on Eq. (59). Based on Eq. (21), we can thus obtain a closed-form solution for momentum SGD. This leads to a different set of constants for the final analysis, but the details are similar. Same for Eq. (48). Thus, we conclude the proof. ■

### D.6. Proof of Equation 7

We show the calculations for Example 1. Recall from Example 1 that

$$G = \frac{1}{2}((w+u)^3 + (w+u')) \text{ and } T = \frac{1}{2}\big((w+u)^3 + (w-u)^3\big),$$

where $u, u' \sim \mathcal{N}(0, \sigma^2)$. Thus, the variance of $G$ is equal to $1/2$ times

$$\begin{aligned}
\mathop{\mathbb{E}}_{u}\left[(w+u)^6\right] - \left(\mathop{\mathbb{E}}_{u}\left[(w+u)^3\right]\right)^2 &= \mathop{\mathbb{E}}_{u}\left[w^6 + 15w^4u^2 + 15w^2u^4 + u^6\right] - \left(\mathop{\mathbb{E}}_{u}\left[w^3 + 3wu^2\right]\right)^2 \\
&= \mathop{\mathbb{E}}_{u}\left[9w^4u^2 + 15w^2u^4 + u^6\right] \\
&= 9w^4\sigma^2 + 45w^2\sigma^4 + 15\sigma^6.
\end{aligned}$$

Besides, the variance of $T$ is equal to $1/2$ times

$$\begin{aligned}
\mathop{\mathbb{E}}_{u}\left[(w^3 + 3wu^2)^2\right] - \left(\mathop{\mathbb{E}}_{u}\left[w^3 + 3wu^2\right]\right)^2 &= \mathop{\mathbb{E}}_{u}\left[9w^2u^4\right] - \left(\mathop{\mathbb{E}}_{u}\left[3wu^2\right]\right)^2 \\
&= 45w^2\sigma^4 - 9w^2\sigma^4 = 36w^2\sigma^4.
\end{aligned}$$

### D.7. Proof of Equation 4

We state Taylor's expansion for equation (4) more formally. Given two matrices $X_1, X_2$ with size $d_1$ by $d_2$, the matrix inner product $\langle X_1, X_2\rangle$ is equal to $\mathrm{Tr}\left[X_1^\top X_2\right]$.

**Proposition 17** *Assume that $f(W)$ is twice-differentiable in $W$. Let $\Sigma$ denote a $d$ by $d$ positive semidefinite matrix (whose eigenvalues are all nonnegative). $U$ denote a sample from $\mathcal{P} = \mathcal{N}(0, \Sigma)$, which has mean zero and covariance $\Sigma \in \mathbb{R}^{d\times d}$. Then, we have that (4) holds.*

**Proof** Given a random perturbation $U$, in expectation of its randomness, we have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{U}\left[\frac{1}{2}\Big(f(W+U) + f(W-U)\Big)\right] \\
&= \frac{1}{2}\left(f(W) + \langle U, \nabla f(W)\rangle + \mathop{\mathbb{E}}_{U}\left[\frac{1}{2}\langle UU^\top, \mathbf{H}\big[f(W)\big]\rangle\right] + \mathrm{O}\left(\|\Sigma\|_2^{3/2}\right)\right) \\
&\quad + \frac{1}{2}\left(f(W) - \langle U, \nabla f(W)\rangle + \mathop{\mathbb{E}}_{U}\left[\frac{1}{2}\langle UU^\top, \mathbf{H}\big[f(W)\big]\rangle\right] + \mathrm{O}\left(\|\Sigma\|_2^{3/2}\right)\right) \\
&= f(W) + \frac{1}{2}\langle \Sigma, \mathbf{H}\big[f(W)\big]\rangle + \mathrm{O}\left(\|\Sigma\|_2^{3/2}\right).
\end{aligned}$$

This completes the proof of equation (4). ■