DIRECTIONAL ANALYSIS OF STOCHASTIC GRADIENT DESCENT VIA VON MISES-FISHER DISTRIBUTIONS IN DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Although stochastic gradient descent (SGD) is a driving force behind the recent success of deep learning, our understanding of its dynamics in a high-dimensional parameter space is limited. In recent years, some researchers have used the stochasticity of minibatch gradients, or the signal-to-noise ratio, to better characterize the learning dynamics of SGD. Inspired from these work, we here analyze SGD from a geometrical perspective by inspecting the stochasticity of the norms and directions of minibatch gradients. We propose a model of the directional concentration for minibatch gradients through von Mises-Fisher (VMF) distribution, and show that the directional uniformity of minibatch gradients increases over the course of SGD. We empirically verify our result using deep convolutional networks and observe a higher correlation between the gradient stochasticity and the proposed directional uniformity than that against the gradient norm stochasticity, suggesting that the directional statistics of minibatch gradients is a major factor behind SGD.

1 INTRODUCTION

Stochastic gradient descent (SGD) has been a driving force behind the recent success of deep learning. Despite a series of work on improving SGD by incorporating the second-order information of the objective function (Roux et al., 2008; Martens, 2010; Dauphin et al., 2014; Martens & Grosse, 2015; Desjardins et al., 2015), SGD is still the most widely used optimization algorithm for training a deep neural network. The learning dynamics of SGD however has not been well characterized beyond that it converges to an extremal point (Bottou, 1998) due to the non-convexity and high-dimensionality of a usual objective function used in deep learning.

Gradient stochasticity, or the signal-to-noise ratio (SNR) of stochastic gradient, has been proposed as a tool for analyzing the learning dynamics of SGD. Shwartz-Ziv & Tishby (2017) identified two phases in SGD based on this. In the first phase, "drift phase", the gradient mean is much higher than its standard deviation, during which optimization progresses rapidly. This drift phase is followed by the "diffusion phase", where SGD behaves similarly to Gaussian noise with very small means. Similar observations were made by Li & Yuan (2017) and Chee & Toulis (2018) who have also divided the learning dynamics of SGD into two phases.

Shwartz-Ziv & Tishby (2017) have proposed that such phase transition is related to information compression. However, Saxe et al. (2018) have reported that the information compression is not generally associated to this transition. Unlike Shwartz-Ziv & Tishby (2017), we notice that there are two aspects to the gradient stochasticity. One is the L^2 norm of the minibatch gradient (the norm stochasticity), and the other is the directional balance of minibatch gradients (the directional stochasticity). SGD converges or terminates when either the norm of the minibatch gradient vanishes to zeros, or when the angles of the minibatch gradients are uniformly distributed and their non-zero norms are close to each other. That is, the gradient stochasticity, or the SNR of stochastic gradient, is driven by both of these aspects, and it is necessary for us to investigate not only the holistic SNR but also the SNR of the minibatch gradient norm and that of the minibatch gradient angles.

In this paper, we use a von Mises-Fisher (VMF) distribution, which is often used in directional statistics (Mardia & Jupp, 2009), and its concentration parameter κ to characterize the directional

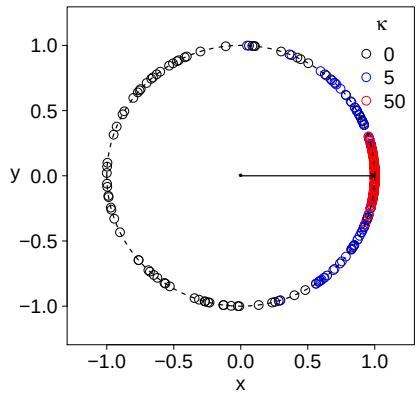


Figure 1: Characteristics of the VMF distribution in a 2-dimensional space. 100 random samples are drawn from VMF$(\boldsymbol{\mu}, \kappa)$ where $\boldsymbol{\mu} = (1, 0)^\top$ and $\kappa = \{0, 5, 50\}$.

balance of minibatch gradients and understand the learning dynamics of SGD from the perspective of directional statistics of minibatch gradients. We prove that SGD increases the directional balance of minibatch gradients. We empirically verify this with deep convolutional networks with various techniques, including batch normalization (Ioffe & Szegedy, 2015) and residual connections (He et al., 2015), on MNIST and CIFAR-10 (Krizhevsky & Hinton, 2009). Our empirical investigation further reveals that the proposed directional stochasticity is a major drive behind the gradient stochasticity compared to the norm stochasticity, suggesting the importance of understanding the directional statistics of stochastic gradient.

2 PRELIMINARIES

Norms and Angles Unless explicitly stated, a norm refers to L^2 norm. $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ thus correspond to L^2 norm and the Euclidean inner product on $\mathbb{R}^d$, respectively. We use $\mathrm{x}_n \Rightarrow \mathrm{x}$ to indicate that "a random variable x_n converges to x in distribution." Similarly, $\mathrm{x}_n \xrightarrow{P} \mathrm{x}$ means convergence in probability. An angle θ between d-dimensional vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined by $\theta = \frac{180}{\pi}\cos^{-1}\left(\frac{\langle \boldsymbol{u}, \boldsymbol{v}\rangle}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|}\right)$.

Loss functions A loss function of a neural network is written as $f(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w})$, where $\boldsymbol{w} \in \mathbb{R}^d$ is a trainable parameter. f_i is "a per-example loss function" computed on the i-th data point. We use $\mathbb{I}$ and m to denote a minibatch index set and its batch size, respectively. Further, we call $f_{\mathbb{I}}(\boldsymbol{w}) = \frac{1}{m}\sum_{i\in\mathbb{I}} f_i(\boldsymbol{w})$ "a minibatch loss function given $\mathbb{I}$". During optimization, we write a parameter $\boldsymbol{w}$ at the i-th iteration in the t-th epoch as $\boldsymbol{w}_t^i$, and $\boldsymbol{w}_0^0$ is an initial parameter. We use n_b to refer to the number of minibatches in a single epoch.

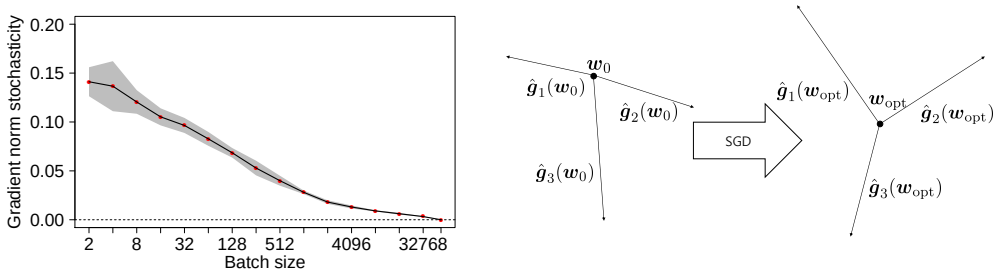
von Mises-Fisher Distribution We use the von Mises-Fisher (VMF) distribution to model the directions of vectors. The definition of the VMF distribution is as follows:

Definition 1. ***(von Mises-Fisher Distribution, Banerjee et al. (2005))*** *The pdf of the* VMF$(\boldsymbol{\mu}, \kappa)$ *is given by*

$$f_d(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) = C_d(\kappa)\exp(\kappa\boldsymbol{\mu}^\top\boldsymbol{x})$$

on the hypersphere $S^{d-1} \subset \mathbb{R}^d$. Here, the concentration parameter κ determines how the samples from this distribution are concentrated on the mean direction $\boldsymbol{\mu}$ and $C_d(\kappa)$ is constant determined by d and κ.

If κ is zero, then it is uniform distribution on a unit hypersphere, and as $\kappa \to \infty$, it becomes a point mass on the unit hypersphere (Figure 1). The maximum likelihood estimates for $\boldsymbol{\mu}$ and κ are



(a) $\sqrt{\mathrm{Var}(\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|)}/\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|$ (b) Directional balance in SGD iterations

Figure 2: The directions of minibatch gradients become important when a batch size m is sufficiently large. (a) The gradient norm stochasticity of $\hat{\boldsymbol{g}}(\boldsymbol{w})$ with respect to various batch sizes at 5 random points $\boldsymbol{w}$ with mean (black line) and mean$\pm$std. (shaded area) in a log-linear scale; (b) If the gradient norm stochasticity is sufficiently low, then the directions of $\hat{\boldsymbol{g}}_i(\boldsymbol{w})$'s must be balanced to satisfy $\sum_{i=1}^{3} \hat{\boldsymbol{g}}_i(\boldsymbol{w}) \approx 0$.

$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i}{\|\sum_{i=1}^{n} \boldsymbol{x}_i\|}$ and $\hat{\kappa} \approx \frac{\bar{r}(d-\bar{r}^2)}{1-\bar{r}^2}$ where $\boldsymbol{x}_i$'s are random samples from the VMF distribution and $\bar{r} = \frac{\|\sum_{i=1}^{n} \boldsymbol{x}_i\|}{n}$ (Banerjee et al., 2005).

3 Theorectical Motivation

3.1 Analysis of the Gradient Norm Stochasticity

It is a usual practice for SGD to use a minibatch gradient $\hat{\boldsymbol{g}}(\boldsymbol{w}) = -\nabla_{\boldsymbol{w}} f_{\mathbb{I}}(\boldsymbol{w})$ instead of a full batch gradient $\boldsymbol{g}(\boldsymbol{w}) = -\nabla_{\boldsymbol{w}} f(\boldsymbol{w})$. The minibatch index set $\mathbb{I}$ is drawn from {1,...,n} randomly. $\hat{\boldsymbol{g}}(\boldsymbol{w})$ satisfies $\mathbb{E}\hat{\boldsymbol{g}}(\boldsymbol{w}) = \boldsymbol{g}(\boldsymbol{w})$ and $\mathrm{Cov}(\hat{\boldsymbol{g}}(\boldsymbol{w}), \hat{\boldsymbol{g}}(\boldsymbol{w})) \approx \frac{1}{mn}\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{w})\boldsymbol{g}_i(\boldsymbol{w})^\top$ for $n \gg m$ where n is the number of full data and $\boldsymbol{g}_i(\boldsymbol{w}) = -\nabla_{\boldsymbol{w}} f_i(\boldsymbol{w})$ (Hoffer et al., 2017). As the batch size m increases, the randomness of $\hat{\boldsymbol{g}}(\boldsymbol{w})$ decreases. Hence $\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|$ tends to $\|\boldsymbol{g}(\boldsymbol{w})\|$, and $\mathrm{Var}(\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|)$, which is the variance of the norm of the minibatch gradient, vanishes. The convergence rate analysis is as the following:

Theorem 1. *Let $\hat{\boldsymbol{g}}(\boldsymbol{w})$ be a minibatch gradient induced from the minibatch index set $\mathbb{I}$ of batch size m from $\{1, \ldots, n\}$ and suppose $\gamma = \max_{i,j\in\{1,\ldots,n\}} |\langle \boldsymbol{g}_i(\boldsymbol{w}), \boldsymbol{g}_j(\boldsymbol{w})\rangle|$. Then*

$$0 \leq \mathbb{E}\big[\|\hat{\boldsymbol{g}}(\boldsymbol{w})\| - \|\boldsymbol{g}(\boldsymbol{w})\|\big] \leq \frac{2(n-m)}{m(n-1)} \times \frac{\gamma}{\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\| + \|\boldsymbol{g}(\boldsymbol{w})\|} \leq \frac{(n-m)\gamma}{m(n-1)\|\boldsymbol{g}(\boldsymbol{w})\|}$$

and

$$\mathrm{Var}(\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|) \leq \frac{2(n-m)}{m(n-1)}\gamma\,.$$

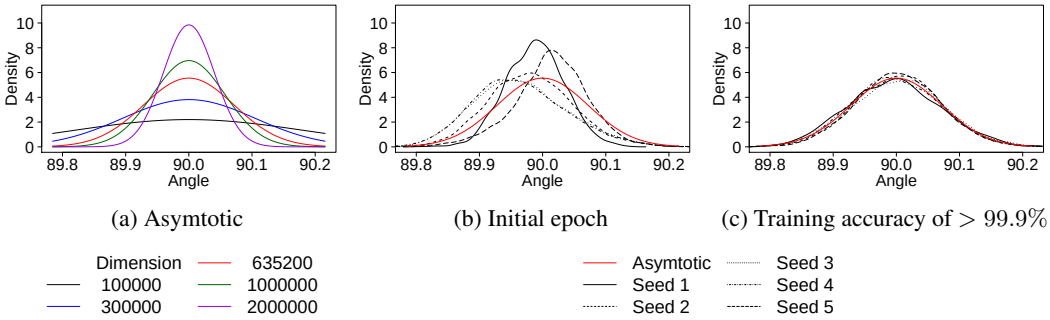
Hence,

$$\frac{\sqrt{\mathrm{Var}(\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|)}}{\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|} \leq \sqrt{\frac{2(n-m)}{m(n-1)} \times \frac{\gamma}{\|\boldsymbol{g}(\boldsymbol{w})\|^2}}\,. \tag{1}$$

Proof. See **Supplemental A**. □

According to **Theorem 1**, the large batch size m reduces the variance of $\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|$ centered at $\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|$ with convergence rate $O(1/m)$. We empirically verify this by estimating the gradient norm stochasticity at various random points while varying the minibatch size, using a fully-connected neural network (**FNN**) with MNIST, as shown in Figure 2(a) (see **Supplemental E** for more details.)

This theorem however only demonstrate that the gradient norm stochasticity is (l.h.s. of (1)) is low at random initial points. It may blow up after SGD updates, since the upper bound (r.h.s. of (1))



(a) Asymtotic (b) Initial epoch (c) Training accuracy of $> 99.9\%$

Figure 3: (a) Asymtotic angle densities (2) of $\theta(\mathbf{u}, \mathbf{v}) = \frac{180}{\pi} \cos^{-1} \langle \mathbf{u}, \mathbf{v} \rangle$ where $\mathbf{u}$ and $\mathbf{v}$ are independent uniformly random unit vectors for each large dimension d. As $d \to \infty$, $\theta(\mathbf{u}, \mathbf{v})$ tends to less scattered from 90 (in degree). (b–c) We apply SGD on **FNN** for MNIST classification with the batch size 64 and the fixed learning rate 0.01 starting from five randomly initialized parameters. We draw a density plot $\theta(\boldsymbol{u}, \frac{\hat{\boldsymbol{g}}_j(\boldsymbol{w}))}{\|\hat{\boldsymbol{g}}_j(\boldsymbol{w}))\|})$ for 3000 minibatch gradients (black) at $\boldsymbol{w} = \boldsymbol{w}_0^0$ (b) and $\boldsymbol{w} = \boldsymbol{w}_{\text{final}}^0$, with training accuracy of $> 99.9\%$, (c) when $\boldsymbol{u}$ is given. After SGD iterations, the density of $\theta(\boldsymbol{u}, \hat{\boldsymbol{g}}_j(\boldsymbol{w}))$ converges to an asymptotic density (red). The dimension of **FNN** is 635,200.

inversely proportional to $\|\boldsymbol{g}(\boldsymbol{w})\|$. This implies that the learning dynamics and convergence of SGD, measured in terms of the vanishing gradient, i.e., $\sum_{i=1}^{n_b} \hat{\boldsymbol{g}}_i(\boldsymbol{w}) \approx 0$, is not necessarily explained by the vanishing norms of minibatch gradients, but rather by the balance of the directions of $\hat{\boldsymbol{g}}_i(\boldsymbol{w})$'s, which motivates our investigation of the directional statistics of minibatch gradients.

3.2 Uniformity measurement via analysis of angles

In order to investigate the directions of minibatch gradients and how they balance, we start from an angle between two vectors. First, we analyze an asymtotic behavior of angles between uniformly random unit vectors in a high-dimensional space.

Theorem 2. *Suppose that* $\mathbf{u}$ *and* $\mathbf{v}$ *are mutually independent* d*-dimensional uniformly random unit vectors. Then,*

$$\sqrt{d}\Big[\frac{180}{\pi} \cos^{-1} \langle \mathbf{u}, \mathbf{v} \rangle - 90\Big] \Rightarrow \mathcal{N}\Big(0, \Big(\frac{180}{\pi}\Big)^2\Big)$$

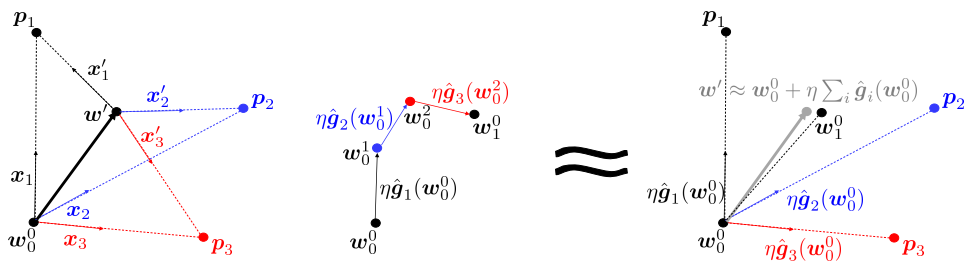
as $d \to \infty$.

Proof. See **Supplemental B**. □

According to **Theorem 2**, the angle between two independent uniformly random unit vectors is normally distributed and becomes incresingly more concentrated as d grows (Figure 3(a)). If SGD iterations indeed drive the directions the directions of minibatch gradients to be uniform, then, at least, the distribution of angles between minibatch gradients and a given uniformly sampled unit vector follows asymptotically

$$\mathcal{N}\Big(90, \Big(\frac{180}{\pi\sqrt{d}}\Big)^2\Big). \tag{2}$$

Figures 3(b) and 3(c) show that the distribution of the angles between minibatch gradients and a given uniformly sampled unit vector converges to an asymptotic distribution (2) after SGD iterations. Although we could measure the uniformity of minibatch gradients how the angle distribution between minibatch gradients is close to (2), it is not as trivial to compare the distributions as to compare numerical values. This necessitates another way to measure the uniformity of minibatch gradients.



(a) From $\boldsymbol{w}_0^0$ to $\boldsymbol{w}'$ by $\epsilon\sum_i \boldsymbol{x}_i$

(b) From $\boldsymbol{w}_0^0$ to $\boldsymbol{w}_1^0$ by SGD iterations

Figure 4: (a) If the point is slightly moved from $\boldsymbol{w}_0^0$ to $\boldsymbol{w}'$ by $\epsilon\sum_i \boldsymbol{x}_i$ where $\boldsymbol{x}_i = (\boldsymbol{p}_i - \boldsymbol{w}_0^0)/\|\boldsymbol{p}_i - \boldsymbol{w}_0^0\|$ and $\boldsymbol{x}'_i = (\boldsymbol{p}_i - \boldsymbol{w}')/\|\boldsymbol{p}_i - \boldsymbol{w}'\|$, then $\|\sum_i \boldsymbol{x}'_i\| < \|\sum_i \boldsymbol{x}_i\|$ which is equivalent to $\kappa(\boldsymbol{w}') < \kappa(\boldsymbol{w}_0^0)$; (b) SGD iterations (left) from $\boldsymbol{w}_0^0$ to $\boldsymbol{w}_1^0$ is approximately equivalent to the circumstance (right) when $\hat{\boldsymbol{g}}_i(\boldsymbol{w}_0^0)$'s are sufficently parallel to $(\boldsymbol{p}_i - \boldsymbol{w}_0^0)$'s for each i.

3.3 Uniformity measurement via VMF distribution

To model the uniformity of minibatch gradients, we propose to use the VMF distribution in **Definition 1**. The concentration parameter κ measures how uniform the directions of unit vectors are distributed. By **Theorem 1**, with a large batch size, the norm of minibatch gradient is nearly deterministic, and $\hat{\boldsymbol{\mu}}$ is almost parallel to the direction of full batch gradient. In other words, κ measures how much the directions of minibatch gradients concentrate around the full batch gradient.

The following **Lemma 1** introduces the relationship between the norm of averaged unit vectors and $\hat{\kappa}$, the estimator of κ.

Lemma 1. *The approximated estimator of κ induced from the d-dimensional unit vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n_b}\}$,*

$$\hat{\kappa} = \frac{\bar{r}(d - \bar{r}^2)}{1 - \bar{r}^2},$$

is a strictly increasing function on $[0, 1]$, where $\bar{r} = \frac{\|\sum_{i=1}^{n_b} \boldsymbol{x}_i\|}{n_b}$. If we consider $\hat{\kappa} = h(u)$ as a function of $u = \|\sum_{i=1}^{n_b} \boldsymbol{x}_i\|$, then $h(\cdot)$ is Lipschitz continuous on $[0, n_b(1-\epsilon)]$ for any $\epsilon > 0$. Moreover, $h(\cdot)$ and $h'(\cdot)$ are strictly increasing and increasing on $[0, n_b)$, respectively.

Proof. See **Supplemental C.1**. □

Consider

$$\hat{\kappa}(\boldsymbol{w}) = h\left(\left\|\sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|}\right\|\right),$$

which is measured from the directions from the current location $\boldsymbol{w}$ to the fixed points $\boldsymbol{p}_i$'s, where $h(\cdot)$ is a function defined in **Lemma 1**. Since $h(\cdot)$ is an increasing function, we may focus only on $\|\sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|}\|$ to see how $\hat{\kappa}$ behaves with respect to its argument. **Lemma 2** implies that the estimated uniformity $\hat{\kappa}$ increases if we move away from $\boldsymbol{w}_0^0$ to $\boldsymbol{w}' = \boldsymbol{w}_0^0 + \epsilon\sum_i \frac{\boldsymbol{p}_i - \boldsymbol{w}_0^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_0^0\|}$ with a small ϵ (Figure 4(a)). In other words, $\hat{\kappa}(\boldsymbol{w}') < \hat{\kappa}(\boldsymbol{w}_0^0)$.

Lemma 2. *Let $\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{n_b}$ be d-dimensional vectors. If all $\boldsymbol{p}_i$'s are not on the ray from the current location $\boldsymbol{w}$, then there exists positive number η such that*

$$\left\|\sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j - \boldsymbol{w} - \epsilon\sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|}}{\left\|\boldsymbol{p}_j - \boldsymbol{w} - \epsilon\sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|}\right\|}\right\| < \left\|\sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|}\right\|$$

for all $\epsilon \in (0, \eta]$.

Proof. See **Supplemental C.2**. □

We make the connection between the observation above and SGD by first viewing $\boldsymbol{p}_i$'s as **local minibatch solutions**.

Definition 2. *For a minibatch index set $\mathbb{I}_i$, $\boldsymbol{p}_i(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}' \in N(\boldsymbol{w}; r_i)} f_{\mathbb{I}_i}(\boldsymbol{w}')$ is a **local minibatch solution** of $\mathbb{I}_i$ at $\boldsymbol{w}$, where $N(\boldsymbol{w}; r_i)$ is a neighborhood of radius r_i at $\boldsymbol{w}$. Here, r_i is determined by $\boldsymbol{w}$ and $\mathbb{I}_i$ for $\boldsymbol{p}_i(\boldsymbol{w})$ to exist uniquely.*

Under this definition, $\boldsymbol{p}_i(\boldsymbol{w})$ is local minimum of a minibatch loss function $f_{\mathbb{I}_i}$ near $\boldsymbol{w}$. Then we reasonably expect that the direction of $\hat{\boldsymbol{g}}_i(\boldsymbol{w}) = -\nabla_{\boldsymbol{w}} f_{\mathbb{I}_i}(\boldsymbol{w})$ is similar to that of $\boldsymbol{p}_i(\boldsymbol{w}) - \boldsymbol{w}$.

Each epoch of SGD with a learning rate η computes a series of $\boldsymbol{w}_t^j = \boldsymbol{w}_t^0 + \eta \sum_{i=1}^{j} \hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})$ for all $j \in \{1, \dots, n_b\}$. If $-\nabla_{\boldsymbol{w}} f(\cdot)$ is Lipschitz continuous, the negative gradient of the i-th iteration in the t-th epoch satisfies $\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1}) \approx \hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)$ for a small η. Moreover, **Theorem 1** implies $\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\| \approx \|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)\| = \tau$ for all $i \in \{1, \dots, n_b\}$ with a large minibatch size or at early stages of SGD iterations.

For example, as in Figure 4(b), suppose that $t = 0$, $n_b = 3$ and $\tau = 1$, and assume that $\boldsymbol{p}_i(\boldsymbol{w}_0^0) = \boldsymbol{p}_i(\boldsymbol{w}_1^0) = \boldsymbol{p}_i$ for all $i = 1, 2, 3$. Then,

$$\hat{\kappa}(\boldsymbol{w}_0^0) = h\left(\left\|\sum_{i=1}^{3} \frac{\boldsymbol{p}_i - \boldsymbol{w}_0^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_0^0\|}\right\|\right), \text{and } \hat{\kappa}(\boldsymbol{w}_1^0) \approx h\left(\left\|\sum_{j=1}^{3} \frac{\boldsymbol{p}_j - \boldsymbol{w}_0^0 - \eta \sum_{i=1}^{3} \hat{\boldsymbol{g}}_i(\boldsymbol{w}_0^{i-1})}{\|\boldsymbol{p}_j - \boldsymbol{w}_0^0 - \eta \sum_{i=1}^{3} \hat{\boldsymbol{g}}_i(\boldsymbol{w}_0^{i-1})\|}\right\|\right).$$

If $\hat{\boldsymbol{g}}_i(\boldsymbol{w}_0^0)$ is parallel to $\boldsymbol{p}_i - \boldsymbol{w}_0^0$ for each i,

$$\sum_{j=1}^{3} \frac{\boldsymbol{p}_j - \boldsymbol{w}_0^0 - \eta \sum_{i=1}^{3} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_0^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_0^{i-1})\|}}{\left\|\boldsymbol{p}_j - \boldsymbol{w}_0^0 - \eta \sum_{i=1}^{3} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_0^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_0^{i-1})\|}\right\|} = \sum_{j=1}^{3} \frac{\boldsymbol{p}_j - \boldsymbol{w}_0^0 - \eta \sum_{i=1}^{3} \frac{\boldsymbol{p}_i - \boldsymbol{w}_0^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_0^0\|}}{\left\|\boldsymbol{p}_j - \boldsymbol{w}_0^0 - \eta \sum_{i=1}^{3} \frac{\boldsymbol{p}_i - \boldsymbol{w}_0^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_0^0\|}\right\|}.$$

Then, by **Lemma 2**, we have $\hat{\kappa}(\boldsymbol{w}_1^0) < \hat{\kappa}(\boldsymbol{w}_0^0)$.

Since $\hat{\boldsymbol{g}}_i(\boldsymbol{w}_0^0)$ may not be perfectly parallel to $\boldsymbol{p}_i - \boldsymbol{w}_0^0$, we further show that a small movement $\eta \sum_i \frac{\boldsymbol{p}_i - \boldsymbol{w}_0^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_0^0\|}$ reducing $\hat{\kappa}$ can be replaced by $\eta \sum_i \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_0^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_0^{i-1})\|}$ in **Theorem 3** below.

Theorem 3. *Let $\boldsymbol{p}_1(\boldsymbol{w}_t^0), \boldsymbol{p}_2(\boldsymbol{w}_t^0), \cdots, \boldsymbol{p}_{n_b}(\boldsymbol{w}_t^0)$ be d-dimensional vectors, and all $\boldsymbol{p}_i(\boldsymbol{w}_t^0)$'s are not on a single ray from the current location $\boldsymbol{w}_t^0$. If*

$$\left\|\sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|} - \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|}\right\| \leq \xi \tag{3}$$

for a sufficiently small $\xi > 0$, then there exists positive number η such that

$$\left\|\sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0 - \epsilon \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|}}{\left\|\boldsymbol{p}_j(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0 - \epsilon \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|}\right\|}\right\| < \left\|\sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|}\right\| \tag{4}$$

for all $\epsilon \in (0, \eta]$.

Proof. See **Supplemental C.3**. □

This **Theorem 3** asserts that $\hat{\kappa}(\cdot)$ decreases even with some perturbation along the averaged direction $\sum_i \frac{\boldsymbol{p}_i(\boldsymbol{w}) - \boldsymbol{w}}{\|\boldsymbol{p}_i(\boldsymbol{w}) - \boldsymbol{w}\|}$. With additional assumptions on each minibatch loss function, we have a sufficient condition for (3), summarized in **Corollary 3.1**.

Corollary 3.1. *Let $\boldsymbol{p}_i$ be the local minibatch solution of each $f_{\mathbb{I}_i}$. Suppose a region $\mathcal{R}$ satisfying:*

$$\textit{For all } \boldsymbol{w}, \boldsymbol{w}' \in \mathcal{R}, \quad \boldsymbol{p}_i(\boldsymbol{w}) = \boldsymbol{p}_i(\boldsymbol{w}') = \boldsymbol{p}_i$$

for all $i = 1, \cdots, n_b$. Further, assume that Hessian matrices of $f_{\mathbb{I}_i}$'s are positive definite, well-conditioned and bounded in the sense of matrix L^2 norm on $\mathcal{R}$. If SGD moves from $\boldsymbol{w}_t^0$ to $\boldsymbol{w}_{t+1}^0$ on $\mathcal{R}$ with a large batch size and a small learning rate, then $\hat{\kappa}(\boldsymbol{w}_t^0) > \hat{\kappa}(\boldsymbol{w}_{t+1}^0)$. Moreover, we can estimate $\hat{\kappa}(\boldsymbol{w}_t^0)$ and $\hat{\kappa}(\boldsymbol{w}_{t+1}^0)$ by minibatch gradients on $\boldsymbol{w}_t^0$ and $\boldsymbol{w}_{t+1}^0$, respectively.

Proof. See **Supplemental D**. □

Without the corollary above, we need to solve $\boldsymbol{p}_i(\boldsymbol{w}_t^0) = \arg\min_{\boldsymbol{w} \in N(\boldsymbol{w}_t^0; r)} f_{\mathbb{I}_i}(\boldsymbol{w})$ for all $i \in \{1, \ldots, n_s\}$, where n_s is the number of samples to estimate κ, in order to compute $\hat{\kappa}(\boldsymbol{w}_t^0)$. **Corollary 3.1** however implies that we can compute $\hat{\kappa}(\boldsymbol{w}_t^0)$ by using $\frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)\|}$ instead of $\frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|}$, significantly reducing computational overhead.

In Practice Although the number of all possible minibatches in each epoch is $n_b = \binom{n}{m}$, it is often the case to use $n_b' \approx n/m$ minibatches at each epoch in practice to go from $\boldsymbol{w}_t^0$ to $\boldsymbol{w}_{t+1}^0$. Assuming that these n_b' minibatches were selected uniformly at random, the average of the n_b' normalized minibatch gradients is the maximum likelihood estimate of $\boldsymbol{\mu}$, just like the average of all n_b normalized minibatch gradients. Thus, we expect with a large n_b',

$$\left\| \sum_{i=1}^{\binom{n}{m}} \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|} - \sum_{i=1}^{n_b'} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|} \right\| \leq \xi,$$

and that SGD in practice also satisfies $\hat{\kappa}(\boldsymbol{w}_t^0) > \hat{\kappa}(\boldsymbol{w}_{t+1}^0)$.

4 Experiments

4.1 Setup

In order to empirically verify our theory on directional statistics of minibatch gradients, we train various types of deep neural networks using SGD and monitor the following metrics for analyzing the learning dynamics of SGD:

- Training loss
- Validation loss
- Gradient stochasticity (GS)$\uparrow$ $\|\mathbb{E}\nabla_{\boldsymbol{w}} f_i(\boldsymbol{w})\|/\mathrm{tr}(\mathrm{Cov}(\nabla_{\boldsymbol{w}} f_i(\boldsymbol{w}), \nabla_{\boldsymbol{w}} f_i(\boldsymbol{w}))$ $\downarrow$
- Gradient norm stochasticity (GNS) $\uparrow$ $\mathbb{E}\|\nabla_{\boldsymbol{w}} f_i(\boldsymbol{w})\|/\sqrt{\mathrm{Var}(\|\nabla_{\boldsymbol{w}} f_i(\boldsymbol{w})\|)}$ $\downarrow$
- Directional Uniformity$\uparrow$ κ $\downarrow$

The latter three quantities are statistically estimated using $n_s = 3,000$ minibatches.

We train the following types of deep neural networks (**Supplemental E**):

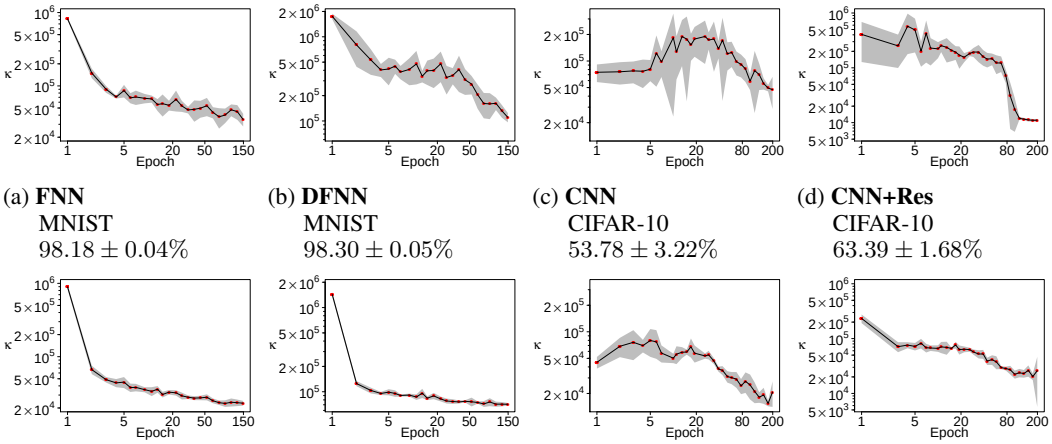
- **FNN**: a fully connected network with a single hidden layer
- **DFNN**: a fully connected network with three hidden layers
- **CNN**: a convolutional network with 14 layers (Krizhevsky et al., 2012)

In the case of the CNN, we also evaluate its variant with skip connections (**+Res**) (He et al., 2015). As it was shown recently by Santurkar et al. (2018) that batch normalization (Ioffe & Szegedy, 2015) improves the smoothness of a loss function in terms of its Hessian, we also test adding batch normalization to each layer right before the ReLU (Nair & Hinton, 2010) nonlinearity (**+BN**). We use MNIST for the FNN, DFNN and their variants, while CIFAR-10 (Krizhevsky & Hinton, 2009) for the CNN and its variants.

We train each model variant using SGD with minibatches of size 64 and a fixed learning rate of 0.01. These were selected so that the training accuracy of $> 99.9\%$. We repeat each setup five times starting from different random initializations and report both the mean and standard deviation.

4.2 Directional Uniformity Increases

FNN and DFNN We first observe that κ decreases over training regardless of the network's depth in Figure 5 (a,b). We however also notice that κ decrease monotonically with the **FNN**, but less so with its deeper variant (**DFNN**). We conjecture this is due to the less-smooth loss landscape of a deep



(a) **FNN** MNIST $98.18 \pm 0.04\%$

(b) **DFNN** MNIST $98.30 \pm 0.05\%$

(c) **CNN** CIFAR-10 $53.78 \pm 3.22\%$

(d) **CNN+Res** CIFAR-10 $63.39 \pm 1.68\%$

(e) **FNN+BN** MNIST $98.37 \pm 0.08\%$

(f) **DFNN+BN** MNIST $98.51 \pm 0.11\%$

(g) **CNN+BN** CIFAR-10 $73.00 \pm 1.16\%$

(h) **CNN+Res+BN** CIFAR-10 $72.71 \pm 1.02\%$

Figure 5: We show the average κ (black curve) $\pm$ std. (shaded area), as the function of the number of training epochs(in log-log scale) for all considered setups. We report the name of architecture, dataset and maximum valid accuracy(mean$\pm$std.) of all epochs. Although κ decreases eventually over time in all the cases, batch normalization (**+BN**) significantly reduces the variance of the directional stochasticity ((a–d) vs. (e–h)). We also observe that the skip connections make κ decrease monotonically ((c,g) vs. (d,h)). Note the differences in the y-scales.

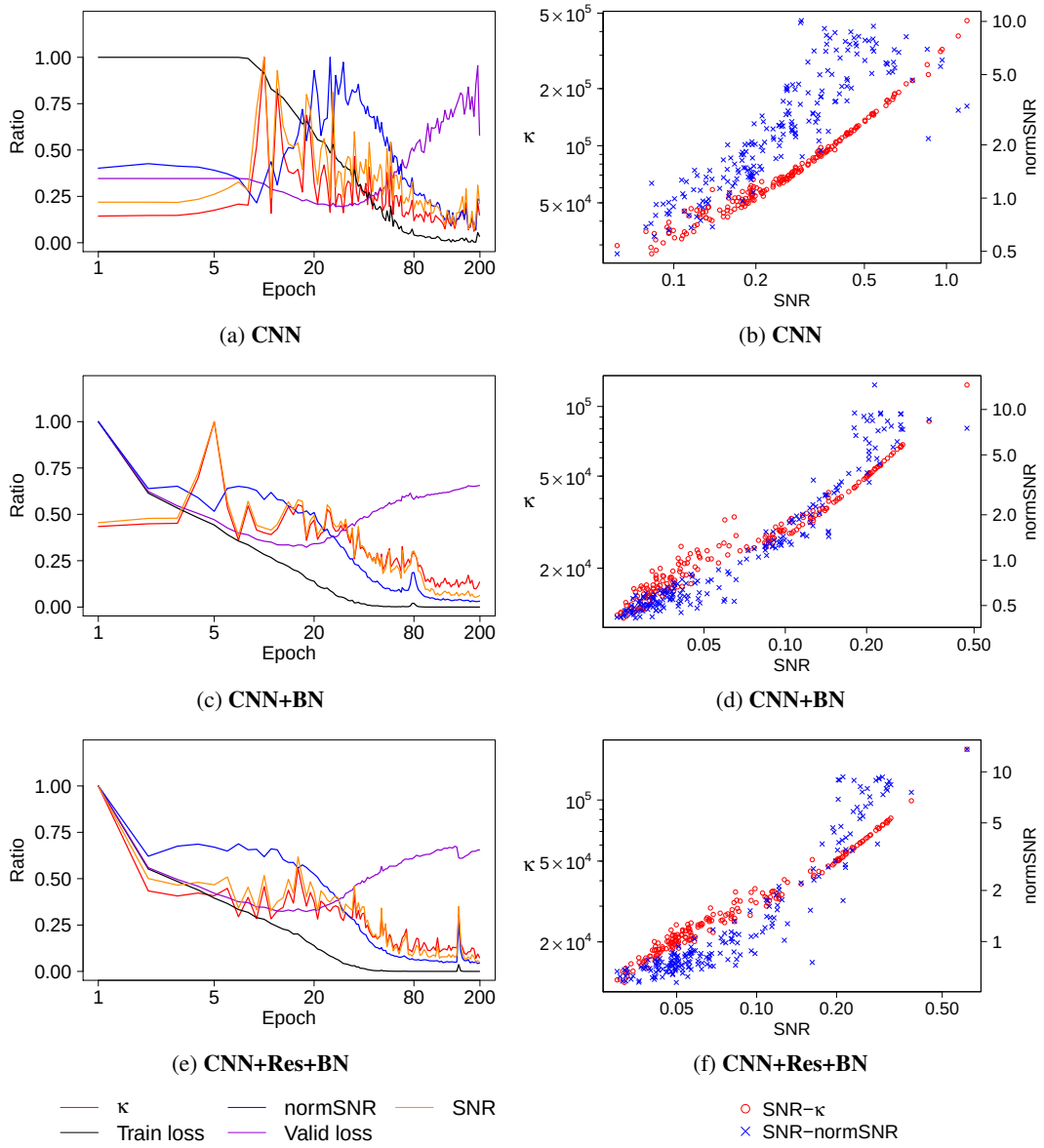
neural network. This difference between **FNN** and **DFNN** however almost entirely vanishes when batch normalization (**+BN**) is applied (Figure 5 (e,f)). This was expected as batch normalization is known to make the loss function behave better, and our theory assumes a smooth objective function.

CNN The **CNN** is substantially deeper than either **FNN** or **DFNN** and is trained on a substantially more difficult problem of CIFAR-10. In other words, the assumptions underlying our theory may not hold as well. Nevetheless, as shown in Figure 5 (c), κ eventually drops below its initial point, although this trend is not monotonic and κ fluctuates significantly over training. The addition of batch normalization (**+BN**) helps with the fluctuation but κ does not monotonically decrease (Figure 5 (g)). On the other hand, we observe the monotonic decrease of κ when skip connections (**+Res**) are introduced (Figure 5 (c) vs. Figure 5 (d)) albeit still with some level of fluctuation especially in the early stage of learning. When both batch normalization and skip connections are used (**+Res+BN**), the behaviour of κ matches with our prediction without much fluctuation.

4.3 Directional Uniformity and Other Metrics

The gradient stochasticity (GS) was used by Shwartz-Ziv & Tishby (2017) as a main metric for identifying two phases of SGD learning in deep neural networks. This quantity includes both the gradient norm stochasticity (GNS) and the directional uniformity κ, implying that either or both of GNS and κ could drive the gradient stochasticity. We thus investigate the relationship among these three quantities as well as training and validation losses. We focus on **CNN**, **CNN+BN** and **CNN+Res+BN** trained on CIFAR-10.

From Figure 6 (a,c,e), it is clear that the proposed metric of directional uniformity κ correlates better with the gradient stochasticity than the gradient norm stochasticity does. This was especially prominent during the early stage of learning, suggesting that the directional statistics of minibatch gradients is a major explanatory factor behind the learning dynamics of SGD. This difference in correlations is much more apparent from the scatterplots in Figure 6 (b,d,f). We show these plots created from other four training runs per setup in **Supplemental F**.



(a) **CNN** (b) **CNN**

(c) **CNN+BN** (d) **CNN+BN**

(e) **CNN+Res+BN** (f) **CNN+Res+BN**

Figure 6: (a,c,e) We plot the evolution of the training loss (Train loss), validation loss (Valid loss), inverse of gradient stochasticity (SNR), inverse of gradient norm stochasticity (normSNR) and directional uniformity κ. We normalized each quantity by its maximum value over training for easier comparison on a single plot. In all the cases, SNR (orange) and κ (red) are almost entirely correlated with each other, while normSNR is less correlated. (b,d,f) We further verify this by illustrating SNR-κ scatter plots (red) and SNR-normSNR scatter plots (blue) in log-log scales. These plots suggest that the SNR is largely driven by the directional uniformity.

5 Conclusion

Stochasticity of gradients is a key to understanding the learning dynamics of SGD (Shwartz-Ziv & Tishby, 2017) and has been pointed out as a factor behind the success of SGD (see, e.g., LeCun et al., 2012; Keskar et al., 2016). In this paper, we provide a theoretical framework using von Mises-Fisher distribution, under which the directional stochasticity of minibatch gradients can be estimated and analyzed, and show that the directional uniformity increases over the course of SGD. Through the extensive empirical evaluation, we have observed that the directional uniformity indeed improves

over the course of training a deep neural network, and that its trend is monotonic when batch normalization and skip connections were used. Furthermore, we demonstrated that the stochasticity of minibatch gradients is largely determined by the directional stochasticity rather than the gradient norm stochasticity.

Our work in this paper suggests two major research directions for the future. First, our analysis has focused on the aspect of optimization, and it is an open question how the directional uniformity relates to the generalization error although handling the stochasticity of gradients has improved SGD (Neelakantan et al., 2015; Hoffer et al., 2017; Smith et al., 2017; Jin et al., 2017). Second, we have focused on passive analysis of SGD using the directional statistics of minibatch gradients, but it is not unreasonable to suspect that SGD could be improved by explicitly taking into account the directional statistics of minibatch gradients during optimization.

REFERENCES

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep): 1345–1382, 2005.

Léon Bottou. Online algorithms and stochastic approximations. In David Saad (ed.), *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998. URL `http://leon.bottou.org/papers/bottou-98x`. revised, oct 2012.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Jerry Chee and Panos Toulis. Convergence diagnostics for stochastic gradient descent with constant learning rate. In *International Conference on Artificial Intelligence and Statistics*, pp. 1476–1485, 2018.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pp. 2933–2941, 2014.

Guillaume Desjardins, Karen Simonyan, Razvan Pascanu, et al. Natural neural networks. In *Advances in Neural Information Processing Systems*, pp. 2071–2079, 2015.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pp. 1729–1739, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

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. *arXiv preprint arXiv:1609.04836*, 2016.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–48. Springer, 2012.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pp. 597–607, 2017.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

Kanti V Mardia and Peter E Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

James Martens. Deep learning via hessian-free optimization. In *ICML*, volume 27, pp. 735–742, 2010.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417, 2015.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

Nicolas L Roux, Pierre-Antoine Manzagol, and Yoshua Bengio. Topmoumoute online natural gradient algorithm. In *Advances in neural information processing systems*, pp. 849–856, 2008.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization?(no, it is not about internal covariate shift). *arXiv preprint arXiv:1805.11604*, 2018.

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. 2018.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Samuel L. Smith, Pieter-Jan Kindermans, and Quoc V. Le. Don't decay the learning rate, increase the batch size. *CoRR*, abs/1711.00489, 2017. URL `http://arxiv.org/abs/1711.00489`.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

SUPPLEMENTARY MATERIAL

A PROOFS FOR THEOREM 1

In proving **Theorem 1**, we use **Lemma A.1**. Define selector random variables(Hoffer et al., 2017) as below:

$$\mathrm{s}_i = \begin{cases} 1, & \text{if } i \in \mathbb{I} \\ 0, & \text{if } i \notin \mathbb{I} \end{cases} .$$

Then we have

$$\hat{\boldsymbol{g}}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{w})\mathrm{s}_i.$$

Lemma A.1. *Let $\hat{\boldsymbol{g}}(\boldsymbol{w})$ be a minibatch gradient induced from the minibatch index set $\mathbb{I}$ with batch size m from $\{1,\dots,n\}$. Then*

$$0 \le \mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|^2 - \|\boldsymbol{g}(\boldsymbol{w})\|^2 \le \frac{2(n-m)}{m(n-1)}\gamma. \tag{5}$$

where $\gamma = \max_{i,j\in\{1,\dots,n\}} |\langle \boldsymbol{g}_i(\boldsymbol{w}), \boldsymbol{g}_j(\boldsymbol{w})\rangle|$.

Proof. By Jensen's inequality, $0 \le \mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|^2 - \|\boldsymbol{g}(\boldsymbol{w})\|^2$. Note that

$$\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{m^2}\langle \boldsymbol{g}_i(\boldsymbol{w}), \boldsymbol{g}_j(\boldsymbol{w})\rangle\, \mathbb{E}[\mathrm{s}_i\mathrm{s}_j].$$

Since $\mathbb{E}[\mathrm{s}_i\mathrm{s}_j] = \frac{m}{n}\delta_{ij} + \frac{m(m-1)}{n(n-1)}(1-\delta_{ij})$,

$$\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|^2 - \|\boldsymbol{g}(\boldsymbol{w})\|^2 &= \Big(\frac{1}{mn} - \frac{m-1}{mn(n-1)}\Big)\sum_{i=1}^{n}\langle \boldsymbol{g}_i(\boldsymbol{w}), \boldsymbol{g}_i(\boldsymbol{w})\rangle \\
&\quad + \Big(\frac{m-1}{mn(n-1)} - \frac{1}{n^2}\Big)\sum_{i=1}^{n}\sum_{j=1}^{n}\langle \boldsymbol{g}_i(\boldsymbol{w}), \boldsymbol{g}_j(\boldsymbol{w})\rangle \\
&= \Big(\frac{1}{mn} - \frac{m-1}{mn(n-1)}\Big)\sum_{i=1}^{n}\langle \boldsymbol{g}_i(\boldsymbol{w}), \boldsymbol{g}_i(\boldsymbol{w})\rangle + \frac{m-n}{mn^2(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle \boldsymbol{g}_i(\boldsymbol{w}), \boldsymbol{g}_j(\boldsymbol{w})\rangle \\
&\le \Big(\frac{1}{mn} - \frac{m-1}{mn(n-1)}\Big)n\gamma + \frac{n-m}{mn^2(n-1)}n^2\gamma \\
&= \frac{2(n-m)}{m(n-1)}\gamma
\end{aligned}$$

where $\gamma = \max_{i,j\in\{1,\dots,n\}} |\langle \boldsymbol{g}_i(\boldsymbol{w}), \boldsymbol{g}_j(\boldsymbol{w})\rangle|$. □

Theorem 1. *Let $\hat{\boldsymbol{g}}(\boldsymbol{w})$ be a minibatch gradient induced from the minibatch index set $\mathbb{I}$ with batch size m from $\{1,\dots,n\}$ and suppose $\gamma = \max_{i,j\in\{1,\dots,n\}} |\langle \boldsymbol{g}_i(\boldsymbol{w}), \boldsymbol{g}_j(\boldsymbol{w})\rangle|$. Then*

$$0 \le \mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\| - \|\boldsymbol{g}(\boldsymbol{w})\| \le \frac{2(n-m)}{m(n-1)} \times \frac{\gamma}{\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\| + \|\boldsymbol{g}(\boldsymbol{w})\|} \le \frac{(n-m)\gamma}{m(n-1)\|\boldsymbol{g}(\boldsymbol{w})\|}$$

and

$$\mathrm{Var}(\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|) \le \frac{2(n-m)}{m(n-1)}\gamma\,.$$

Hence,

$$\frac{\sqrt{\mathrm{Var}(\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|)}}{\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|} \le \sqrt{\frac{2(n-m)}{m(n-1)} \times \frac{\gamma}{\|\boldsymbol{g}(\boldsymbol{w})\|^2}}\,.$$

Proof. By Jensen's inequality, we have $\|\boldsymbol{g}(\boldsymbol{w})\| = \|\mathbb{E}\hat{\boldsymbol{g}}(\boldsymbol{w})\| \leq \mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|$ and $\{\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|\}^2 \leq \mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|^2$. From the second inequality and **Lemma A.1**,

$$\{\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|\}^2 \leq \mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|^2 \leq \|\boldsymbol{g}(\boldsymbol{w})\|^2 + \frac{2(n-m)}{m(n-1)}\gamma$$

or

$$\big(\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\| - \|\boldsymbol{g}(\boldsymbol{w})\|\big)\big(\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\| + \|\boldsymbol{g}(\boldsymbol{w})\|\big) \leq \frac{2(n-m)}{m(n-1)}\gamma\,.$$

Hence

$$\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\| \leq \|\boldsymbol{g}(\boldsymbol{w})\| + \frac{2(n-m)}{m(n-1)} \times \frac{\gamma}{\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\| + \|\boldsymbol{g}(\boldsymbol{w})\|} \leq \frac{(n-m)\gamma}{m(n-1)\|\boldsymbol{g}(\boldsymbol{w})\|}.$$

Further,

$$\begin{aligned}\mathrm{Var}(\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|) &= \mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|^2 - \{\mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|\}^2 \\ &\leq \mathbb{E}\|\hat{\boldsymbol{g}}(\boldsymbol{w})\|^2 - \|\mathbb{E}\hat{\boldsymbol{g}}(\boldsymbol{w})\|^2 \\ &\leq \|\boldsymbol{g}(\boldsymbol{w})\|^2 + \frac{2(n-m)}{m(n-1)}\gamma - \|\boldsymbol{g}(\boldsymbol{w})\|^2 = \frac{2(n-m)}{m(n-1)}\gamma.\end{aligned}$$

□

B PROOFS FOR THEOREM 2

For our proofs, Slutsky's theorem and delta method are key results to describe limiting behaviors of random variables in distribution sense.

Theorem B.1. ***(Slutsky's theorem, Casella & Berger (2002))*** *Let $\{\mathrm{x}_n\}$, $\{\mathrm{y}_n\}$ be a sequence of random variables that satisfies $\mathrm{x}_n \Rightarrow \mathrm{x}$ and $\mathrm{y}_n \xrightarrow{P} \rho$ when n goes to infinity and ρ is constant. Then*

$$\mathrm{x}_n\mathrm{y}_n \Rightarrow c\mathrm{x}$$

Theorem B.2. ***(Delta method, Casella & Berger (2002))*** *Let y_n be a sequence of random variables that satisfies $\sqrt{n}(\mathrm{y}_n - \mu) \Rightarrow \mathcal{N}(0, \sigma^2)$. For a given smooth function $f : \mathbb{R} \to \mathbb{R}$, suppose that $f'(\mu)$ exists and is not 0 where f' is a derivative. Then*

$$\sqrt{n}[f(\mathrm{y}_n) - f(\mu)] \Rightarrow \mathcal{N}(0, \sigma^2[f'(\mu)]^2).$$

Lemma B.1. *Suppose that $\mathbf{u}$ and $\mathbf{v}$ are mutually independent d-dimensional uniformly random unit vectors. Then, $\sqrt{d}\,\langle \mathbf{u}, \mathbf{v}\rangle \Rightarrow \mathcal{N}(0,1)$ as $d \to \infty$.*

Proof. Note that d-dimensional uniformly random unit vectors $\mathbf{u}$ can be generated by normalization of d-dimensional multivariate standard normal random vectors $\mathbf{x} \sim N(\mathbf{0}, \boldsymbol{I}_d)$. That is,

$$\mathbf{u} \sim \frac{\mathbf{x}}{\|\mathbf{x}\|}.$$

Suppose that two independent uniformly random unit vector $\mathbf{u}$ and $\mathbf{v}$ are generated by two independent d-dimensional standard normal vector $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_d)$ and $\mathbf{y} = (\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_d)$. Denote them

$$\mathbf{u} = \frac{\mathbf{x}}{\|\mathbf{x}\|} \quad \text{and} \quad \mathbf{v} = \frac{\mathbf{y}}{\|\mathbf{y}\|}.$$

By SLLN, we have

$$\frac{\|\mathbf{x}\|}{\sqrt{d}} \to 1 \quad a.s.$$

(Use $\frac{1}{d}\sum_{i=1}^{d} \mathrm{x}_i^2 \to \mathbb{E}[\mathrm{x}_1^2] = 1$). Since almost sure convergence implies convergence in probability, $\|\mathbf{x}\|/\sqrt{d} \xrightarrow{P} 1$. Similarly, $\|\mathbf{y}\|/\sqrt{d} \xrightarrow{P} 1$. Moreover, by CLT,

$$\frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\sqrt{d}} = \sqrt{d}\Big(\frac{1}{d}\sum_{i=1}^{d} \mathrm{x}_i\mathrm{y}_i\Big) \Rightarrow \mathcal{N}(0,1).$$

Therefore, by **Theorem B.1** (Slutsky's theorem),

$$\sqrt{d}\,\langle \mathbf{u}, \mathbf{v}\rangle \Rightarrow \mathcal{N}(0,1).$$

□

Theorem 2. *Suppose that* $\mathbf{u}$ *and* $\mathbf{v}$ *are mutually independent* d*-dimensional uniformly random unit vectors. Then,*

$$\sqrt{d}\Big[\frac{180}{\pi}\cos^{-1}\langle \mathbf{u},\mathbf{v}\rangle - 90\Big] \Rightarrow \mathcal{N}\Big(0,\Big(\frac{180}{\pi}\Big)^2\Big)$$

as $d \to \infty$.

Proof. Suppose that $\mu = 0$, $\sigma = 1$, and $f(\cdot) = \frac{180}{\pi}\cos^{-1}(\cdot)$. Since $\frac{180}{\pi}\frac{d}{dx}\cos^{-1}(x) = -\frac{180}{\pi\sqrt{1-x^2}}$, we have $f'(\mu) = -\frac{180}{\pi}$. Hence, by **Lemma B.1** and **Theorem B.2** (Delta method),

$$\sqrt{d}\Big[\frac{180}{\pi}\cos^{-1}\langle \mathbf{u},\mathbf{v}\rangle - 90\Big] \Rightarrow \mathcal{N}\Big(0,\Big(\frac{180}{\pi}\Big)^2\Big).$$

□

C PROOFS FOR THEOREM 3

C.1 PROOF OF LEMMA 1

Lemma 1. *The approximated estimator of* κ *induced from the* d*-dimensional unit vectors* $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n_b}\}$,

$$\hat{\kappa} = \frac{\bar{r}(d-\bar{r}^2)}{1-\bar{r}^2},$$

where $\bar{r} = \frac{\|\sum_{i=1}^{n_b}\boldsymbol{x}_i\|}{n_b}$ *is a strict increasing function on* $[0,1]$. *We can also consider* $\hat{\kappa} = h(u)$ *as function of* $u = \|\sum_{i=1}^{n_b}\boldsymbol{x}_i\|$. *Then* $h(\cdot)$ *is Lipschitz continuous on* $[0, n_b(1-\epsilon)]$ *for any* $\epsilon > 0$. *Moreover,* $h(\cdot)$ *and* $h'(\cdot)$ *are strict increasing and increasing on* $[0, n_b)$*, respectively.*

Proof. Note that $\|\sum_{i=1}^{n_b}\boldsymbol{x}_i\| \le \sum_{i=1}^{n_b}||\boldsymbol{x}_i|| = n_b$. Therefore, we have $\bar{r} \in [0,1]$. If $d = 1$, then $\hat{\kappa} = \bar{r}$ and this increases on $[0,1]$. For $d > 1$,

$$\frac{d\hat{\kappa}}{d\bar{r}} = \frac{d + \bar{r}^4 + (d - 3\bar{r}^2)}{(1-\bar{r}^2)^2}$$

and its numerator is always positive for $d > 2$. When $d = 2$,

$$\frac{d\hat{\kappa}}{d\bar{r}} = \frac{\bar{r}^4 - 3\bar{r}^2 + 4}{(1-\bar{r}^2)^2} = \frac{(\bar{r}^2 - \frac{3}{2})^2 + \frac{7}{4}}{(1-\bar{r}^2)^2} > 0.$$

So $\hat{\kappa}$ increases as $\bar{r}$ increases.

The Lipschitz continuity of $h(\cdot)$ directly comes from the continuity of $\frac{d\hat{\kappa}}{d\bar{r}}$ since

$$\frac{d\hat{\kappa}}{du} = \frac{1}{n_b}\frac{d\hat{\kappa}}{d\bar{r}}.$$

Recall that any continuous function on the compact interval $[0, n_b(1-\epsilon)]$ is bounded. Hence the derivative of $\hat{\kappa}$ with respect to u is bounded. This implies the Lipschitz continuity of $h(\cdot)$.

$h(\cdot)$ is strictly increasing since $\bar{r} = \frac{u}{n_b}$. Further,

$$\begin{aligned} h''(u) &= \frac{1}{{n_b}^2}\frac{d^2\hat{\kappa}}{d\bar{r}^2} \\ &= \frac{2\bar{r}^5 + (4-8d)\bar{r}^3 + (8d-6)\bar{r}}{{n_b}^2(1-\bar{r}^2)^4} > 0 \end{aligned}$$

due to $\bar{r} \in [0,1]$. Therefore $h'(\cdot)$ is also increasing on $[0, n_b)$. □

C.2 PROOF OF LEMMA 2

Lemma 2. *Let $\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{n_b}$ be d-dimensional vectors. If all $\boldsymbol{p}_i$'s are not on the ray from the current location $\boldsymbol{w}$, then there exists positive number η such that*

$$\Big\| \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j - \boldsymbol{w} - \epsilon \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|}}{\|\boldsymbol{p}_j - \boldsymbol{w} - \epsilon \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|}\|} \Big\| < \Big\| \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|} \Big\|$$

for all $\epsilon \in (0, \eta]$.

Proof. Without loss of generality, we regard $\boldsymbol{w}$ as the origin. Let $f(\epsilon) = \Big\| \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j - \epsilon \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i}{\|\boldsymbol{p}_i\|}}{\|\boldsymbol{p}_j - \epsilon \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i}{\|\boldsymbol{p}_i\|}\|} \Big\|^2$, then $f(0) = \Big\| \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i}{\|\boldsymbol{p}_i\|} \Big\|^2$. Therefore, we only need to show $f'(0) < 0$. Now denote $\boldsymbol{x}_j = \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|}$, $\boldsymbol{p}_j(\epsilon) = \boldsymbol{p}_j - \epsilon \sum_{i=1}^{n_b} \boldsymbol{x}_i$ and $\boldsymbol{u} = -\sum_{i=1}^{n_b} \boldsymbol{x}_i$. That is, $\boldsymbol{p}_j(\epsilon) = \boldsymbol{p}_j + \epsilon \boldsymbol{u}$. Since

$$f(\epsilon) = \Big\langle \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j(\epsilon)}{\|\boldsymbol{p}_j(\epsilon)\|}, \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j(\epsilon)}{\|\boldsymbol{p}_j(\epsilon)\|} \Big\rangle,$$

we have

$$f'(\epsilon) = 2\Big\langle \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j(\epsilon)}{\|\boldsymbol{p}_j(\epsilon)\|}, \frac{d}{d\epsilon}\Big(\sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j(\epsilon)}{\|\boldsymbol{p}_j(\epsilon)\|} \Big) \Big\rangle$$

and

$$\frac{d}{d\epsilon}\Big(\sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j(\epsilon)}{\|\boldsymbol{p}_j(\epsilon)\|} \Big) = \sum_{j=1}^{n_b} \frac{\|\boldsymbol{p}_j(\epsilon)\| \boldsymbol{u} - \frac{\langle \boldsymbol{u}, \boldsymbol{p}_j(\epsilon) \rangle}{\|\boldsymbol{p}_j(\epsilon)\|} \boldsymbol{p}_j(\epsilon)}{\|\boldsymbol{p}_j(\epsilon)\|^2}.$$

Hence

$$f'(\epsilon) = 2\Big\langle \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j(\epsilon)}{\|\boldsymbol{p}_j(\epsilon)\|}, \sum_{j=1}^{n_b} \frac{\|\boldsymbol{p}_j(\epsilon)\| \boldsymbol{u} - \frac{\langle \boldsymbol{u}, \boldsymbol{p}_j(\epsilon) \rangle}{\|\boldsymbol{p}_j(\epsilon)\|} \boldsymbol{p}_j(\epsilon)}{\|\boldsymbol{p}_j(\epsilon)\|^2} \Big\rangle.$$

Note that $\boldsymbol{p}_j(0) = \boldsymbol{p}_j$ and $\|\boldsymbol{x}_j\| = 1$. We have

$$\begin{aligned}
f'(0) &= 2\Big\langle \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|}, \sum_{j=1}^{n_b} \frac{\|\boldsymbol{p}_j\| \boldsymbol{u} - \frac{\langle \boldsymbol{u}, \boldsymbol{p}_j \rangle}{\|\boldsymbol{p}_j\|} \boldsymbol{p}_j}{\|\boldsymbol{p}_j\|^2} \Big\rangle \\
&= 2\Big\langle \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|}, \sum_{j=1}^{n_b} \frac{1}{\|\boldsymbol{p}_j\|} \Big(\boldsymbol{u} - \Big\langle \boldsymbol{u}, \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|} \Big\rangle \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|} \Big) \Big\rangle \\
&= 2\Big\langle \sum_{j=1}^{n_b} \boldsymbol{x}_j, \sum_{j=1}^{n_b} \frac{1}{\|\boldsymbol{p}_j\|} \Big(\boldsymbol{u} - \langle \boldsymbol{u}, \boldsymbol{x}_j \rangle \boldsymbol{x}_j \Big) \Big\rangle \\
&= 2\Big\langle -\boldsymbol{u}, \sum_{j=1}^{n_b} \frac{1}{\|\boldsymbol{p}_j\|} \Big(\boldsymbol{u} - \langle \boldsymbol{u}, \boldsymbol{x}_j \rangle \boldsymbol{x}_j \Big) \Big\rangle \\
&= -2 \sum_{j=1}^{n_b} \frac{\|\boldsymbol{u}\|^2 - \langle \boldsymbol{u}, \boldsymbol{x}_j \rangle^2}{\|\boldsymbol{p}_j\|} \\
&\leq -2 \sum_{j=1}^{n_b} \frac{\|\boldsymbol{u}\|^2 - \|\boldsymbol{u}\|^2 \|\boldsymbol{x}_j\|^2}{\|\boldsymbol{p}_j\|} \\
&= 0
\end{aligned}$$

Since the equality holds when $\langle \boldsymbol{u}, \boldsymbol{x}_j \rangle^2 = \|\boldsymbol{u}\|^2 \|\boldsymbol{x}_j\|^2$ for all j, we can have strict inequality when all $\boldsymbol{p}_i$'s are not on the same ray from the origin. □

C.3 PROOF OF THEOREM 3

The proof of **Theorem 3** is very similar to **Lemma 2**.

Theorem 3. *Let $\boldsymbol{p}_1(\boldsymbol{w}_t^0), \boldsymbol{p}_2(\boldsymbol{w}_t^0), \cdots, \boldsymbol{p}_{n_b}(\boldsymbol{w}_t^0)$ be d-dimensional vectors, and all $\boldsymbol{p}_i(\boldsymbol{w}_t^0)$'s are not on the ray from the current location $\boldsymbol{w}_t^0$. If*

$$\Big\| \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|} - \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|} \Big\| \leq \xi \tag{6}$$

for sufficiently small $\xi > 0$, then there exists positive number η such that

$$\Big\| \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0 - \epsilon \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|}}{\|\boldsymbol{p}_j(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0 - \epsilon \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|}\|} \Big\| < \Big\| \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|} \Big\| \tag{7}$$

for all $\epsilon \in (0, \eta]$.

Proof. Similarly, we regard $\boldsymbol{w}_t^0$ as the origin $\mathbf{0}$. For simplicity, write $\boldsymbol{p}_i(\mathbf{0})$ and $\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})$ as $\boldsymbol{p}_i$ and $\hat{\boldsymbol{g}}_i$, respectively. Let $f(\epsilon) = \Big\| \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j - \epsilon \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i}{\|\boldsymbol{p}_i\|}}{\|\boldsymbol{p}_j - \epsilon \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i}{\|\boldsymbol{p}_i\|}\|} \Big\|^2$ and $\tilde{f}(\epsilon) = \Big\| \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j - \epsilon \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i}{\|\hat{\boldsymbol{g}}_i\|}}{\|\boldsymbol{p}_j - \epsilon \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i}{\|\hat{\boldsymbol{g}}_i\|}\|} \Big\|^2$. Denote $\boldsymbol{u} = -\sum_{j=1}^{n_b} \frac{\boldsymbol{p}_i}{\|\boldsymbol{p}_i\|}$, $\boldsymbol{t} = \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i}{\|\boldsymbol{p}_i\|} - \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i}{\|\hat{\boldsymbol{g}}_i\|}$ and $\tilde{\boldsymbol{p}}_j(\epsilon) = \boldsymbol{p}_j + \epsilon(\boldsymbol{u} + \boldsymbol{t})$. Then

$$\tilde{f}(\epsilon) = \Big\| \sum_{i=1}^{n_b} \frac{\tilde{\boldsymbol{p}}_j(\epsilon)}{\|\tilde{\boldsymbol{p}}_j(\epsilon)\|} \Big\|^2.$$

Now we differentiate $\tilde{f}(\epsilon)$ with respect to ϵ, that is,

$$\tilde{f}'(\epsilon) = 2\Big\langle \sum_{j=1}^{n_b} \frac{\tilde{\boldsymbol{p}}_j(\epsilon)}{\|\tilde{\boldsymbol{p}}_j(\epsilon)\|}, \sum_{j=1}^{n_b} \frac{\|\tilde{\boldsymbol{p}}_j(\epsilon)\|(\boldsymbol{u} + \boldsymbol{t}) - \frac{\langle \boldsymbol{u}+\boldsymbol{t}, \tilde{\boldsymbol{p}}_j(\epsilon)\rangle}{\|\tilde{\boldsymbol{p}}_j(\epsilon)\|} \tilde{\boldsymbol{p}}_j(\epsilon)}{\|\tilde{\boldsymbol{p}}_j(\epsilon)\|^2} \Big\rangle.$$

Recall that $\tilde{\boldsymbol{p}}_j(0) = \boldsymbol{p}_j$. Rewrite $\frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|} = \boldsymbol{x}_j$ and use $f'(0)$ in the proof of **Lemma 2**

$$\begin{aligned}
\tilde{f}'(0) &= 2\Big\langle \sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|}, \sum_{j=1}^{n_b} \frac{\|\boldsymbol{p}_j\|(\boldsymbol{u} + \boldsymbol{t}) - \frac{\langle \boldsymbol{u}+\boldsymbol{t}, \boldsymbol{p}_j\rangle}{\|\boldsymbol{p}_j\|} \boldsymbol{p}_j}{\|\boldsymbol{p}_j\|^2} \Big\rangle \\
&= 2\Big\langle -\boldsymbol{u}, \sum_{j=1}^{n_b} \frac{\boldsymbol{u} + \boldsymbol{t} - \big\langle \boldsymbol{u} + \boldsymbol{t}, \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|} \big\rangle \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|}}{\|\boldsymbol{p}_j\|} \Big\rangle \\
&= 2\Big\langle -\boldsymbol{u}, \sum_{j=1}^{n_b} \frac{1}{\|\boldsymbol{p}_j\|} \Big(\boldsymbol{u} + \boldsymbol{t} - \langle \boldsymbol{u} + \boldsymbol{t}, \boldsymbol{x}_j \rangle \boldsymbol{x}_j \Big) \Big\rangle \\
&= 2\Big\langle -\boldsymbol{u}, \sum_{j=1}^{n_b} \frac{1}{\|\boldsymbol{p}_j\|} \Big(\boldsymbol{u} - \langle \boldsymbol{u}, \boldsymbol{x}_j \rangle \boldsymbol{x}_j \Big) \Big\rangle + 2\Big\langle -\boldsymbol{u}, \sum_{j=1}^{n_b} \frac{1}{\|\boldsymbol{p}_j\|} \Big(\boldsymbol{t} - \langle \boldsymbol{t}, \boldsymbol{x}_j \rangle \boldsymbol{x}_j \Big) \Big\rangle \\
&= f'(0) - 2\sum_{j=1}^{n_b} \frac{1}{\|\boldsymbol{p}_j\|} \big(\langle \boldsymbol{u}, \boldsymbol{t} \rangle - \langle \boldsymbol{t}, \boldsymbol{x}_j \rangle \langle \boldsymbol{u}, \boldsymbol{x}_j \rangle \big)
\end{aligned}$$

Since $f'(0) < 0$ by the proof of **Lemma 2**,

$$\tilde{f}'(0) < 0 \quad \Longleftrightarrow \quad 2\sum_{j=1}^{n_b} \frac{1}{\|\boldsymbol{p}_j\|} \big(\langle \boldsymbol{t}, \boldsymbol{x}_j \rangle \langle \boldsymbol{u}, \boldsymbol{x}_j \rangle - \langle \boldsymbol{u}, \boldsymbol{t} \rangle \big) < |f'(0)|.$$

By using $\|\boldsymbol{x}_j\| = 1$ and applying the Cauchy inequality,

$$\begin{aligned}
2\sum_{j=1}^{n_b} \frac{1}{\|\boldsymbol{p}_j\|}\big(\langle \boldsymbol{t}, \boldsymbol{x}_j\rangle \langle \boldsymbol{u}, \boldsymbol{x}_j\rangle - \langle \boldsymbol{u}, \boldsymbol{t}\rangle\big) &\leq 2\sum_{j=1}^{n_b} \frac{\|\boldsymbol{t}\|\|\boldsymbol{x}_j\|\|\boldsymbol{u}\|\|\boldsymbol{x}_j\| + \|\boldsymbol{u}\|\|\boldsymbol{t}\|}{\|\boldsymbol{p}_j\|} \\
&= 4\sum_{j=1}^{n_b} \frac{\|\boldsymbol{u}\|\|\boldsymbol{t}\|}{\|\boldsymbol{p}_j\|} \\
&\leq \frac{4n_b\|\boldsymbol{u}\|\|\boldsymbol{t}\|}{\min_j \|\boldsymbol{p}_j\|} \\
&\leq \frac{4n_b^2\xi}{\min_j \|\boldsymbol{p}_j\|} \quad \Big(\because \|\boldsymbol{u}\| \leq \sum_{i=1}^{n_b} \|\boldsymbol{x}_i\| = n_b\Big).
\end{aligned}$$

Define $r = \min_j \|\boldsymbol{p}_j\|$. If

$$\xi < \frac{|f'(0)|r}{4n_b^2},$$

then

$$\tilde{f}(\epsilon) < 0.$$

□

D PROOFS FOR COROLLARY 3.1

Corollary 3.1. *Let $\boldsymbol{p}_i$'s be local minibatch solutions for each $f_{\mathbb{I}_i}$'s. Suppose that a region $\mathcal{R}$ satisfies:*

$$\text{For all } \boldsymbol{w}, \boldsymbol{w}' \in \mathcal{R}, \quad \boldsymbol{p}_i(\boldsymbol{w}) = \boldsymbol{p}_i(\boldsymbol{w}') = \boldsymbol{p}_i$$

for all $i = 1, \cdots, n_b$. Further, assume that Hessian matrices of $f_{\mathbb{I}_i}$'s are positive definite, well-conditioned and bounded in the sense of matrix L^2 norm on $\mathcal{R}$. If SGD moves $\boldsymbol{w}_t^0$ to $\boldsymbol{w}_{t+1}^0$ on $\mathcal{R}$ with a large batch size and a small learning rate, then $\hat{\kappa}(\boldsymbol{w}_t^0) > \hat{\kappa}(\boldsymbol{w}_{t+1}^0)$. Moreover, we can estimate $\hat{\kappa}(\boldsymbol{w}_t^0)$ and $\hat{\kappa}(\boldsymbol{w}_{t+1}^0)$ by minibatch gradients on $\boldsymbol{w}_t^0$ and $\boldsymbol{w}_{t+1}^0$, respectively.

Proof. Recall that $\boldsymbol{w}_{t+1}^0 = \boldsymbol{w}_t^0 + \eta \sum_{i=1}^{n_b} \hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})$ where η is a learning rate. To prove **Corollary 3.1**, we need to show $\hat{\kappa}(\boldsymbol{w}_{t+1}^0) < \hat{\kappa}(\boldsymbol{w}_t^0)$ which is equivalent to

$$\Big\|\sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j(\boldsymbol{w}_{t+1}^0) - \boldsymbol{w}_t^0 - \eta\sum_{i=1}^{n_b} \hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\boldsymbol{p}_j(\boldsymbol{w}_{t+1}^0) - \boldsymbol{w}_t^0 - \eta\sum_{i=1}^{n_b} \hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|}\Big\| < \Big\|\sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|}\Big\|. \tag{8}$$

Since $\|\nabla_{\boldsymbol{w}}{}^2 f_{\mathbb{I}_i}(\cdot)\|_2$ is bounded on $\mathcal{R}$, $\nabla_{\boldsymbol{w}} f_{\mathbb{I}_i}(\cdot)$ is Lipschitz continuous on $\mathcal{R}$(Bottou, 2010). If the batch size is sufficiently large and the learning rate η is sufficiently small, $\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\| \approx \|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)\| \approx \tau$ for all i by **Theorem 1**. Therefore, we have

$$\eta\sum_{i=1}^{n_b} \hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1}) \approx \tau\eta \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|}$$

If we denote $\tau\eta$ as ϵ, we can convert (8) to (9).

$$\Big\|\sum_{j=1}^{n_b} \frac{\boldsymbol{p}_j(\boldsymbol{w}_{t+1}^0) - \boldsymbol{w}_t^0 - \epsilon\sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|}}{\|\boldsymbol{p}_j(\boldsymbol{w}_{t+1}^0) - \boldsymbol{w}_t^0 - \epsilon\sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|}\|}\Big\| < \Big\|\sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|}\Big\|. \tag{9}$$

Since both $\boldsymbol{w}_{t+1}^0$ and $\boldsymbol{w}_t^0$ are in $\mathcal{R}$ for small learning rate, we have $\boldsymbol{p}_i(\boldsymbol{w}_{t+1}^0) = \boldsymbol{p}_i(\boldsymbol{w}_t^0) = \boldsymbol{p}_i$ by the assumption. That is, (9) is equivalent to (7). In (7), $\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})/\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|$ cannot be replaced by $(\boldsymbol{p}_i - \boldsymbol{w}_t^0)/\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|$ in general. Hence we introduce **Definition D.1** and **Lemma D.1** to connect the direction of the minibatch gradient with the corresponding local minibatch solution.

Definition D.1. *The condition number $c(\boldsymbol{A})$ of a matrix $\boldsymbol{A}$ is defined as*

$$c(\boldsymbol{A}) = \frac{\sigma_{\max}(\boldsymbol{A})}{\sigma_{\min}(\boldsymbol{A})}$$

where $\sigma_{\max}(\boldsymbol{A})$ and $\sigma_{\min}(\boldsymbol{A})$ are maximal and minimal singular values of $\boldsymbol{A}$, respectively. If $\boldsymbol{A}$ is positive-definite matrix, then

$$c(\boldsymbol{A}) = \frac{\lambda_{\max}(\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{A})}.$$

Here $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ are maximal and minimal eigenvalues of $\boldsymbol{A}$, respectively.

Lemma D.1. *If the condition number of the positive definite Hessian matrix of $f_{\mathbb{I}_i}$ at the local minibatch solution, $\boldsymbol{p}_i$, denoted by $\boldsymbol{H}_i = \nabla_{\boldsymbol{w}}{}^2 f_{\mathbb{I}_i}(\boldsymbol{p}_i)$ is approximately 1(well-conditioned), then the direction to $\boldsymbol{p}_i$ from $\boldsymbol{w}$ is almost parallel to its negative gradient at $\boldsymbol{w}$. That is, for all $\boldsymbol{w} \in \mathcal{R}$,*

$$\Big\| \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|} - \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w})\|} \Big\| \approx 0$$

where $\hat{\boldsymbol{g}}_i(\boldsymbol{w}) = -\nabla_{\boldsymbol{w}} f_{\mathbb{I}_i}(\boldsymbol{w})$.

Proof. By the second order Taylor expansion,

$$f_{\mathbb{I}_i}(\boldsymbol{w}) \approx f_{\mathbb{I}_i}(\boldsymbol{p}_i) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{p}_i)^\top \boldsymbol{H}_i(\boldsymbol{w} - \boldsymbol{p}_i).$$

Hence,

$$\hat{\boldsymbol{g}}_i(\boldsymbol{w}) = -\nabla_{\boldsymbol{w}} f_{\mathbb{I}_i}(\boldsymbol{w}) \approx -\boldsymbol{H}_i(\boldsymbol{w} - \boldsymbol{p}_i)$$

Denote $\boldsymbol{p}_i - \boldsymbol{w}$ as $\boldsymbol{x}$. Then, we only need to show

$$\Big\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \frac{\boldsymbol{H}_i\boldsymbol{x}}{\|\boldsymbol{H}_i\boldsymbol{x}\|} \Big\|^2 \approx 0$$

Note that $\boldsymbol{H}_i$ is positive definite, so we can diagonalize it as $\boldsymbol{H}_i = \boldsymbol{P}_i^\top \boldsymbol{\Lambda}_i \boldsymbol{P}_i$ where $\boldsymbol{P}_i$ is an orthonormal transition matrix for $\boldsymbol{H}_i$.

$$\begin{aligned}
\Big\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \frac{\boldsymbol{H}_i\boldsymbol{x}}{\|\boldsymbol{H}_i\boldsymbol{x}\|} \Big\|^2 &= 2 - 2\frac{\boldsymbol{x}^\top \boldsymbol{H}_i \boldsymbol{x}}{\|\boldsymbol{x}\|\|\boldsymbol{H}_i\boldsymbol{x}\|} \\
&= 2 - 2\frac{(\boldsymbol{P}_i\boldsymbol{x})^\top \boldsymbol{\Lambda}_i \boldsymbol{P}_i \boldsymbol{x}}{\|\boldsymbol{P}_i\boldsymbol{x}\|\|\boldsymbol{P}_i^\top \boldsymbol{\Lambda}_i \boldsymbol{P}_i \boldsymbol{x}\|} \\
&= 2 - 2\frac{(\boldsymbol{P}_i\boldsymbol{x})^\top \boldsymbol{\Lambda}_i \boldsymbol{P}_i \boldsymbol{x}}{\|\boldsymbol{P}_i\boldsymbol{x}\|\|\boldsymbol{\Lambda}_i \boldsymbol{P}\boldsymbol{x}\|} \\
&\leq 2 - 2\frac{\sum_j \lambda_j (\boldsymbol{P}_i\boldsymbol{x})_j^2}{\|\boldsymbol{P}_i\boldsymbol{x}\|\|\boldsymbol{\Lambda}_i \boldsymbol{P}_i \boldsymbol{x}\|} \\
&\leq 2 - 2\frac{\lambda_{\min}\|\boldsymbol{P}_i\boldsymbol{x}\|^2}{\|\boldsymbol{P}_i\boldsymbol{x}\|\lambda_{\max}\|\boldsymbol{P}_i\boldsymbol{x}\|} \\
&= 2 - 2\frac{\lambda_{\min}}{\lambda_{\max}} \approx 0
\end{aligned}$$

□

Lemma D.1 proposed that well-conditioned Hessian matrix of $f_{\mathbb{I}_i}$ at $\boldsymbol{p}_i$ makes $\hat{\boldsymbol{g}}_i(\boldsymbol{w})/\|\hat{\boldsymbol{g}}_i(\boldsymbol{w})\|$ be replaced by $(\boldsymbol{p}_i - \boldsymbol{w})/\|\boldsymbol{p}_i - \boldsymbol{w}\|$ for all $\boldsymbol{w} \in \mathcal{R}$. Using this, we prove **Lemma D.2**.

Lemma D.2. *Let $\boldsymbol{w}$ be a parameter in $\mathcal{R}$. If the condition number of Hessian matrix of $f_{\mathbb{I}_i}$ is sufficiently near 1(well-conditioned) and $\frac{\|\boldsymbol{w} - \boldsymbol{w}_t^0\|}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|}$ is sufficiently near 0, then*

$$\Big\| \frac{\boldsymbol{p}_i - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|} - \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w})\|} \Big\| \leq \frac{\xi}{n_b}$$

for sufficiently small ξ.

Proof. We have

$$\begin{aligned}
\Big\| \frac{\boldsymbol{p}_i - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|} - \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w})\|} \Big\| &\leq \Big\| \frac{\boldsymbol{p}_i - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|} - \frac{\tilde{\boldsymbol{g}}(\boldsymbol{w})}{\|\tilde{\boldsymbol{g}}(\boldsymbol{w})\|} \Big\| + \Big\| \frac{\tilde{\boldsymbol{g}}(\boldsymbol{w})}{\|\tilde{\boldsymbol{g}}(\boldsymbol{w})\|} - \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w})\|} \Big\| \\
&\leq \Big\| \frac{\boldsymbol{p}_i - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|} - \frac{\tilde{\boldsymbol{g}}(\boldsymbol{w})}{\|\tilde{\boldsymbol{g}}(\boldsymbol{w})\|} \Big\| + \Big\| \frac{\tilde{\boldsymbol{g}}(\boldsymbol{w})}{\|\tilde{\boldsymbol{g}}(\boldsymbol{w})\|} - \frac{\boldsymbol{p}_i - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|} \Big\| \\
&\quad + \Big\| \frac{\boldsymbol{p}_i - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|} - \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|} \Big\| + \Big\| \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|} - \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w})\|} \Big\| \\
&\leq \epsilon + \epsilon + \sqrt{2\Big(1 - \frac{\langle \boldsymbol{p}_i - \boldsymbol{w}_t^0, \boldsymbol{p}_i - \boldsymbol{w} \rangle}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\| \|\boldsymbol{p}_i - \boldsymbol{w}\|}\Big)} + \epsilon \\
&= 3\epsilon + \sqrt{2\Big(1 - \frac{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|^2 - \langle \boldsymbol{p}_i - \boldsymbol{w}_t^0, \boldsymbol{w} - \boldsymbol{w}_t^0 \rangle}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\| \|\boldsymbol{p}_i - \boldsymbol{w}\|}\Big)}
\end{aligned}$$

for sufficently small ϵ(See **Lemma D.1**). Now we only need to show

$$\sqrt{2\Big(1 - \frac{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|^2 - \langle \boldsymbol{p}_i - \boldsymbol{w}_t^0, \boldsymbol{w} - \boldsymbol{w}_t^0 \rangle}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\| \|\boldsymbol{p}_i - \boldsymbol{w}\|}\Big)} < \epsilon \tag{10}$$

Since $\frac{\|\boldsymbol{w} - \boldsymbol{w}_t^0\|}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|}$ is sufficiently small, we have

$$\begin{aligned}
\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|^2 - \big\langle \boldsymbol{p}_i - \boldsymbol{w}_t^0, \boldsymbol{w} - \boldsymbol{w}_t^0 \big\rangle &= \|\boldsymbol{p}_i - \boldsymbol{w}_t^0\| \Big(\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\| - \Big\langle \frac{\boldsymbol{p}_i - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|}, \boldsymbol{w} - \boldsymbol{w}_t^0 \Big\rangle \Big) \\
&\geq \|\boldsymbol{p}_i - \boldsymbol{w}_t^0\| (\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\| - \|\boldsymbol{w} - \boldsymbol{w}_t^0\|) \\
&= \|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|^2 \Big(1 - \frac{\|\boldsymbol{w} - \boldsymbol{w}_t^0\|}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|}\Big) \\
&\geq 0.
\end{aligned}$$

By using the above non-negativeness, we have the following inequality.

$$\begin{aligned}
1 &\geq \frac{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|^2 - \big\langle \boldsymbol{p}_i - \boldsymbol{w}_t^0, \boldsymbol{w} - \boldsymbol{w}_t^0 \big\rangle}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\| \|\boldsymbol{p}_i - \boldsymbol{w}\|} \\
&\geq \frac{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|^2 - \big\langle \boldsymbol{p}_i - \boldsymbol{w}_t^0, \boldsymbol{w} - \boldsymbol{w}_t^0 \big\rangle}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\| (\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\| + \|\boldsymbol{w} - \boldsymbol{w}_t^0\|)} \\
&= \frac{\frac{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|}{\|\boldsymbol{w} - \boldsymbol{w}_t^0\|} - \big\langle \frac{\boldsymbol{p}_i - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|}, \frac{\boldsymbol{w} - \boldsymbol{w}_t^0}{\|\boldsymbol{w} - \boldsymbol{w}_t^0\|} \big\rangle}{1 + \frac{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|}{\|\boldsymbol{w} - \boldsymbol{w}_t^0\|}} \\
&\geq \frac{\frac{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|}{\|\boldsymbol{w} - \boldsymbol{w}_t^0\|} - 1}{1 + \frac{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|}{\|\boldsymbol{w} - \boldsymbol{w}_t^0\|}}.
\end{aligned} \tag{11}$$

As $\frac{\|\boldsymbol{w} - \boldsymbol{w}_t^0\|}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|} \to 0^+$, (11) is monotonically increasing to 1. This implies that (10) holds for sufficiently small ϵ. □

With small learning rate, $\boldsymbol{w}_t^{i-1}$'s are in $\mathcal{R}$ for all $i \in \{1, \ldots, n_b\}$. As a result, by **Lemma D.2**, we have

$$\Big\| \frac{\boldsymbol{p}_i - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i - \boldsymbol{w}_t^0\|} - \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|} \Big\| \leq \frac{\xi}{n_b} \tag{12}$$

for sufficiently small ξ. This implies (6) since

$$\Big\| \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|} - \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|} \Big\| \leq \sum_{i=1}^{n_b} \Big\| \frac{\boldsymbol{p}_i - \boldsymbol{w}}{\|\boldsymbol{p}_i - \boldsymbol{w}\|} - \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^{i-1})\|} \Big\|.$$

Then we can apply **Theorem 3** and $\hat{\kappa}(\boldsymbol{w}_t^0) > \hat{\kappa}(\boldsymbol{w}_{t+1}^0)$ holds.

For the last statement, *"Moreover, we can estimate $\hat{\kappa}(\boldsymbol{w}_t^0)$ and $\hat{\kappa}(\boldsymbol{w}_{t+1}^0)$ by minibatch gradients on $\boldsymbol{w}_t^0$ and $\boldsymbol{w}_{t+1}^0$, respectively."*, recall that

$$\hat{\kappa}(\boldsymbol{w}_t^0) = h\Big(\Big\| \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|} \Big\|\Big)$$

where $h(\cdot)$ is increasing and Lipschitz continuous(**Lemma 1**). By **Lemma D.1**, we have

$$\Big\| \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|} - \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)\|} \Big\| < \frac{\xi}{n_b}$$

for sufficiently small $\xi > 0$. Therefore,

$$\Big| \Big\| \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|} \Big\| - \Big\| \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)\|} \Big\| \Big| \leq \Big\| \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|} - \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)\|} \Big\|$$

where rhs is bounded by ξ. Hence, Lipschitz continuity of $h(\cdot)$ implies that

$$\Big| h\Big(\Big\| \sum_{i=1}^{n_b} \frac{\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0}{\|\boldsymbol{p}_i(\boldsymbol{w}_t^0) - \boldsymbol{w}_t^0\|} \Big\|\Big) - h\Big(\Big\| \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)\|} \Big\|\Big)\Big| \to 0$$

as $\xi \to 0$. That is,

$$\hat{\kappa}(\boldsymbol{w}_t^0) \approx h\Big(\Big\| \sum_{i=1}^{n_b} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)}{\|\hat{\boldsymbol{g}}_i(\boldsymbol{w}_t^0)\|} \Big\|\Big)\Big).$$

Since t is arbitrary, we can apply this for all other $\boldsymbol{w} \in \mathcal{R}$ including $\boldsymbol{w}_{t+1}^0$. □

E Experimental Details

E.1 Model Architecture

For all cases, their weighted layers do not have biases, and dropout (Srivastava et al., 2014) is not applied. We use Xavier initializations(Glorot & Bengio, 2010) and cross entropy loss functions for all experiments.

FNN The **FNN** is a fully connected network with a single hidden layer. It has 800 hidden units with ReLU (Nair & Hinton, 2010) activations and a softmax output layer.

DFNN The **DFNN** is a fully connected network with three hidden layers. It has 800 hidden units with ReLU activations in each hidden layers and a softmax output layer.

CNN The network architecture of **CNN** is introduced in He et al. (2016) as a CIFAR-10 plain network. The first layer is 3×3 convolution layer and the number of output filters are 16. After that, we stack of {4, 4, 4} layers with 3×3 convolutions on the feature maps of sizes {32, 16, 8} and the numbers of filters {16, 32, 64}, respectively. The subsampling is performed with a stride of 2. All convolution layers are activated by ReLU and the convolution part ends with a global average pooling(Lin et al., 2013), a 10-way fully-conneted layers, and softmax. Note that there are 14 stacked weighted layers.

+BN We apply batch normalization right before the ReLU activations on all hidden layers.

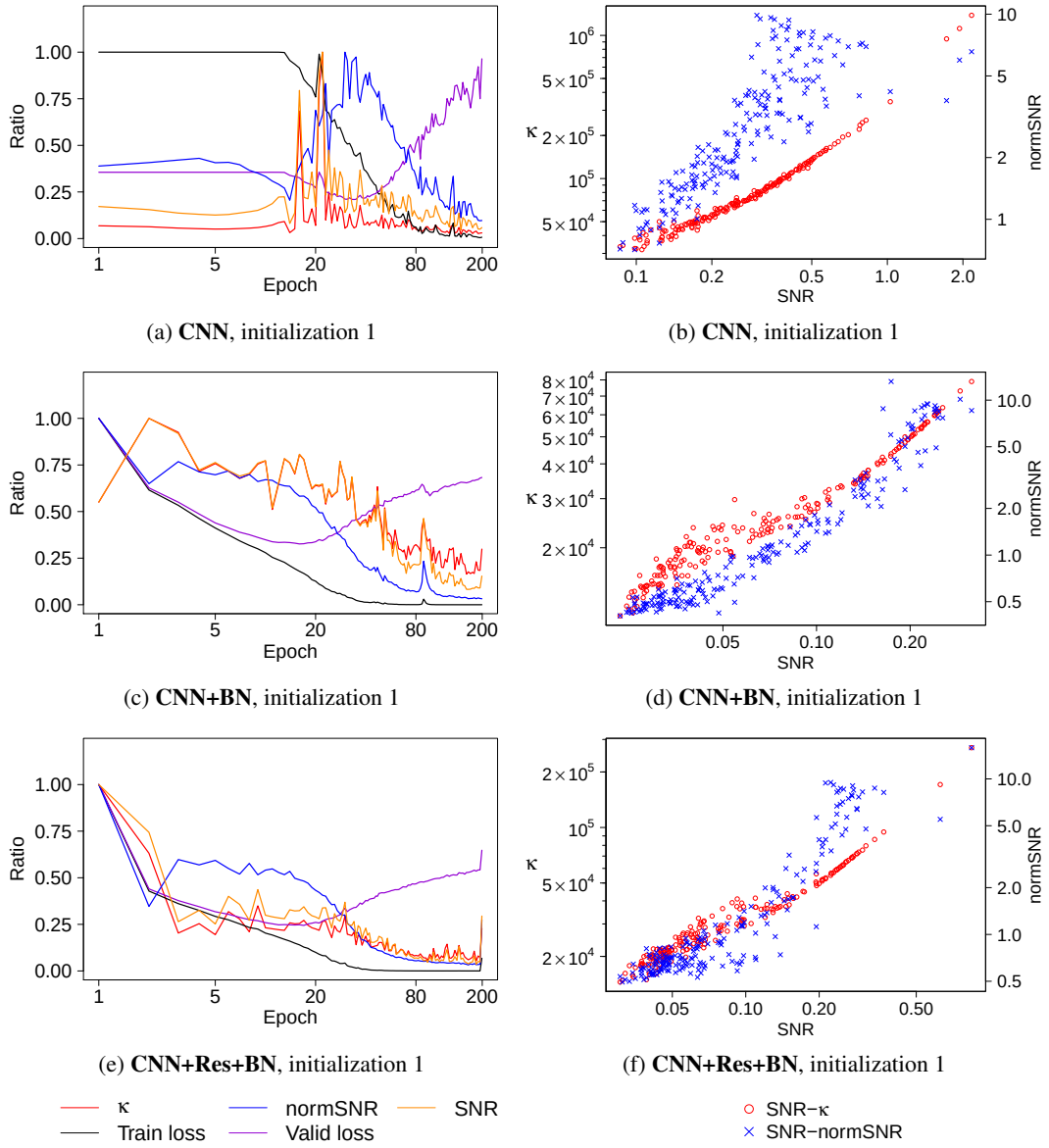
+Res The identity skip connections are added after every two convolution layers before ReLU nonlinearity (After batch normalization, if it is applied on it.).

E.2 Data Augmentation

We use neither data augmentations nor preprocessings except scaling pixel values into $[0, 1]$ both MNIST and CIFAR-10.

F OTHER FOUR TRAINING RUNS IN FIGURE 6

We show plots from other four training runs in Figure 6. For all runs, the curves of GS (inverse of SNR) and κ are strongly correlated while GNS (inverse of normSNR) is less correlated to GS.



(a) **CNN**, initialization 1

(b) **CNN**, initialization 1

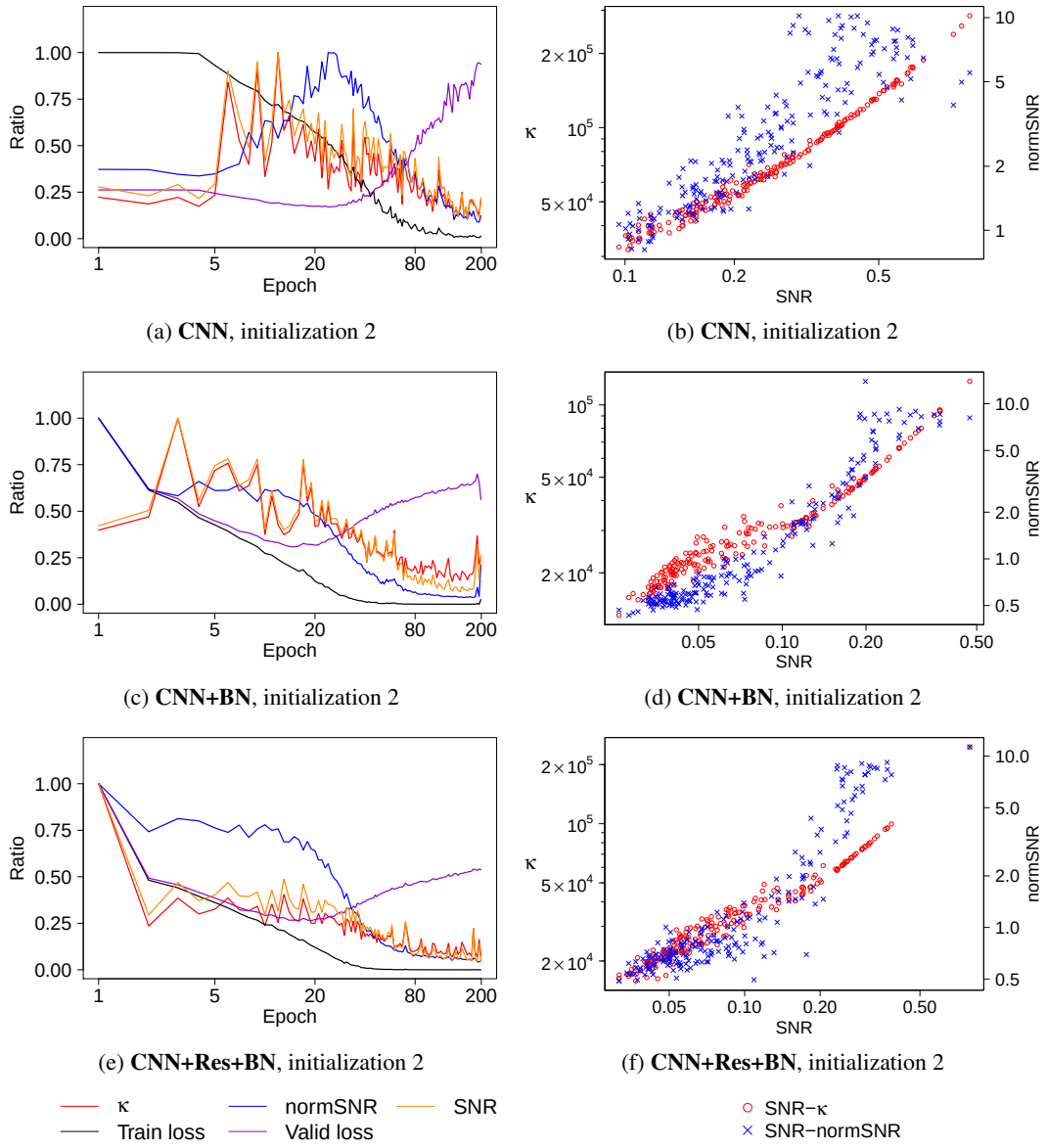
(c) **CNN+BN**, initialization 1

(d) **CNN+BN**, initialization 1

(e) **CNN+Res+BN**, initialization 1

(f) **CNN+Res+BN**, initialization 1

Figure 7: (a,c,e) We plot the evolution of the training loss (Train loss), validation loss (Valid loss), inverse of gradient stochasticity (SNR), inverse of gradient norm stochasticity (normSNR) and directional uniformity κ. We normalized each quantity by its maximum value over training for easier comparison on a single plot. In all the cases, SNR (orange) and κ (red) are almost entirely correlated with each other, while normSNR is less correlated. (b,d,f) We further verify this by illustrating SNR-κ scatter plots (red) and SNR-normSNR scatter plots (blue) in log-log scales. These plots suggest that the SNR is largely driven by the directional uniformity.



(a) **CNN**, initialization 2

(b) **CNN**, initialization 2

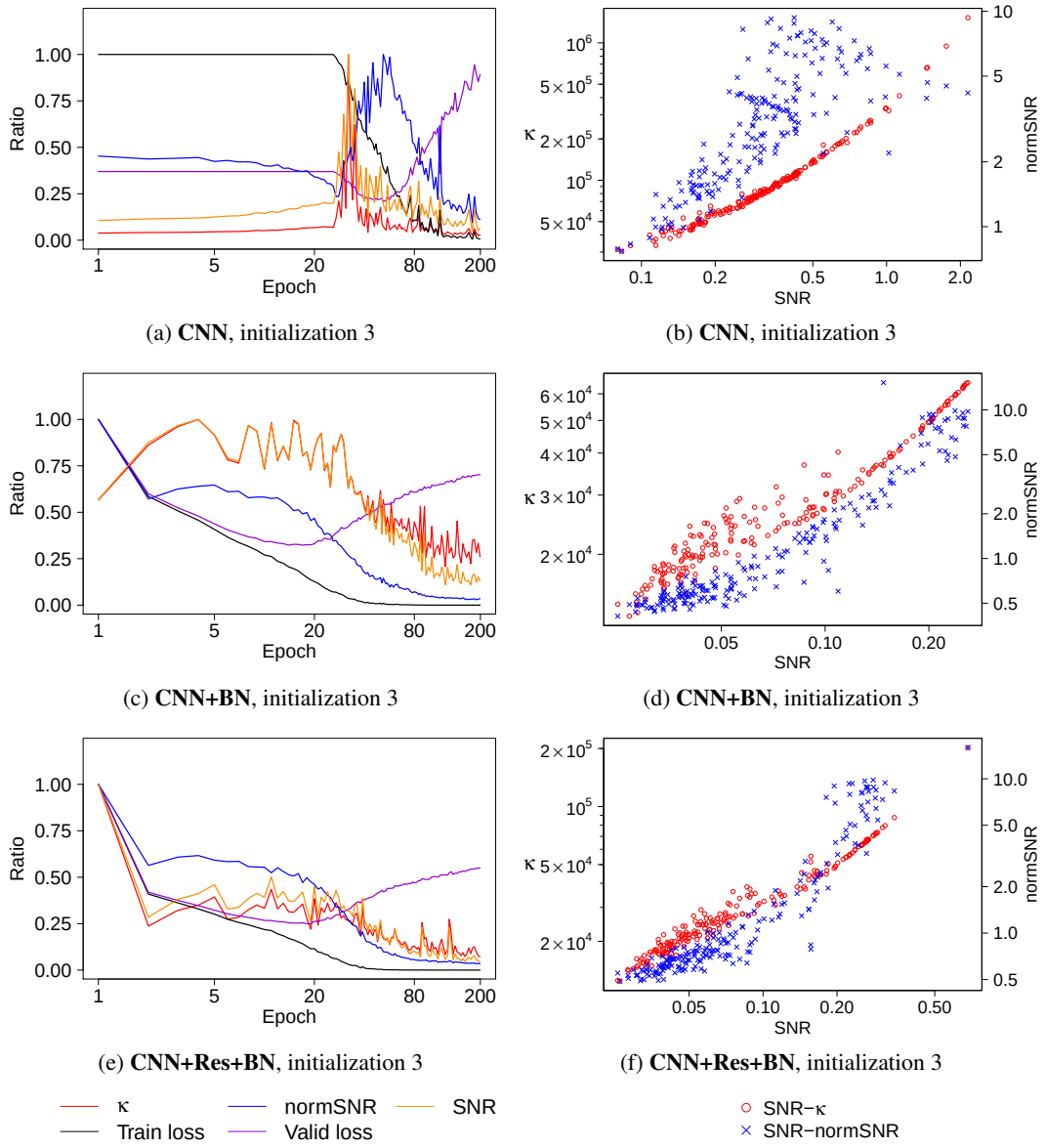
(c) **CNN+BN**, initialization 2

(d) **CNN+BN**, initialization 2

(e) **CNN+Res+BN**, initialization 2

(f) **CNN+Res+BN**, initialization 2

Figure 8: (a,c,e) We plot the evolution of the training loss (Train loss), validation loss (Valid loss), inverse of gradient stochasticity (SNR), inverse of gradient norm stochasticity (normSNR) and directional uniformity κ. We normalized each quantity by its maximum value over training for easier comparison on a single plot. In all the cases, SNR (orange) and κ (red) are almost entirely correlated with each other, while normSNR is less correlated. (b,d,f) We further verify this by illustrating SNR-κ scatter plots (red) and SNR-normSNR scatter plots (blue) in log-log scales. These plots suggest that the SNR is largely driven by the directional uniformity.



(a) **CNN**, initialization 3

(b) **CNN**, initialization 3

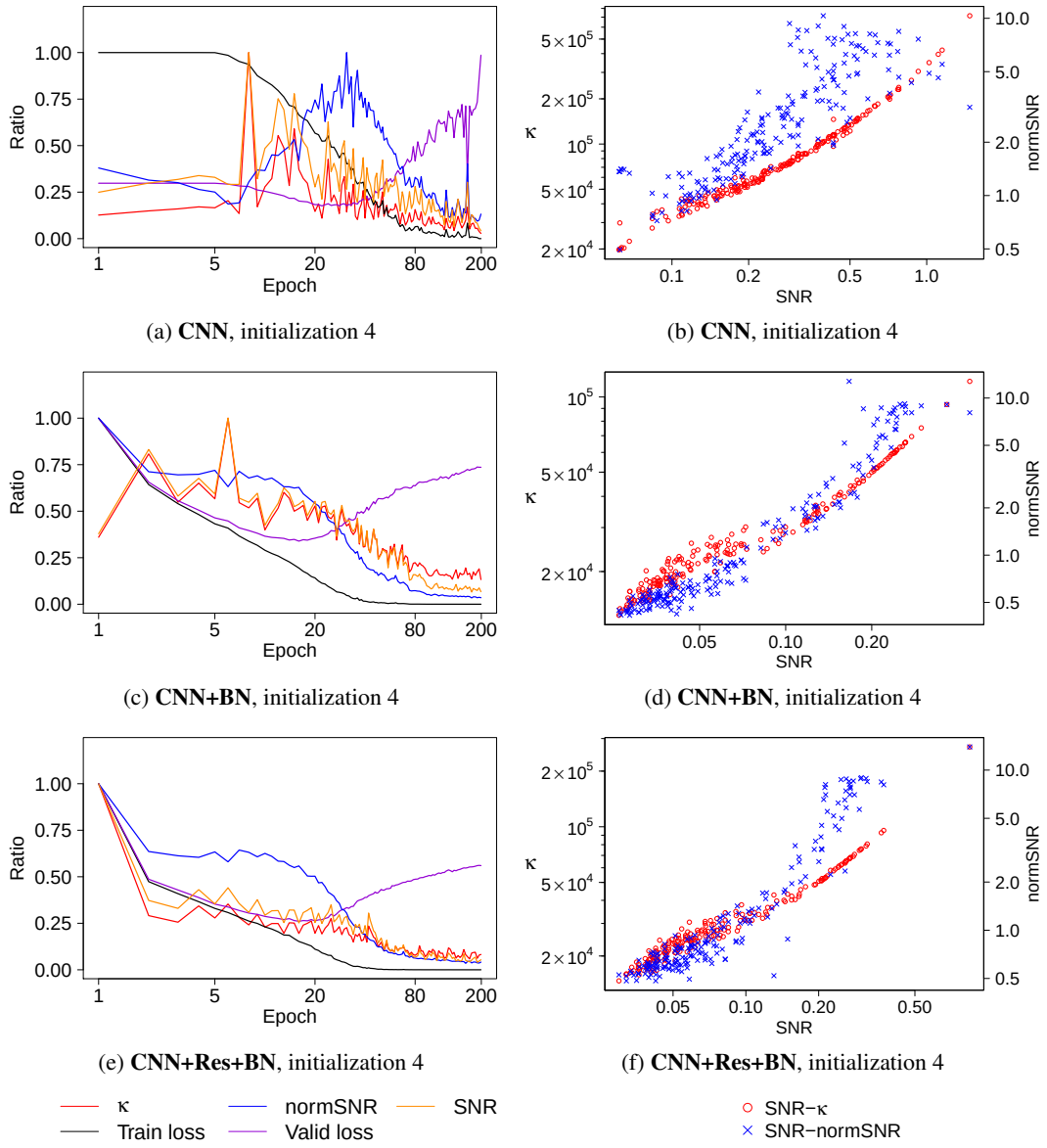
(c) **CNN+BN**, initialization 3

(d) **CNN+BN**, initialization 3

(e) **CNN+Res+BN**, initialization 3

(f) **CNN+Res+BN**, initialization 3

Figure 9: (a,c,e) We plot the evolution of the training loss (Train loss), validation loss (Valid loss), inverse of gradient stochasticity (SNR), inverse of gradient norm stochasticity (normSNR) and directional uniformity κ. We normalized each quantity by its maximum value over training for easier comparison on a single plot. In all the cases, SNR (orange) and κ (red) are almost entirely correlated with each other, while normSNR is less correlated. (b,d,f) We further verify this by illustrating SNR-κ scatter plots (red) and SNR-normSNR scatter plots (blue) in log-log scales. These plots suggest that the SNR is largely driven by the directional uniformity.



(a) **CNN**, initialization 4

(b) **CNN**, initialization 4

(c) **CNN+BN**, initialization 4

(d) **CNN+BN**, initialization 4

(e) **CNN+Res+BN**, initialization 4

(f) **CNN+Res+BN**, initialization 4

Figure 10: (a,c,e) We plot the evolution of the training loss (Train loss), validation loss (Valid loss), inverse of gradient stochasticity (SNR), inverse of gradient norm stochasticity (normSNR) and directional uniformity κ. We normalized each quantity by its maximum value over training for easier comparison on a single plot. In all the cases, SNR (orange) and κ (red) are almost entirely correlated with each other, while normSNR is less correlated. (b,d,f) We further verify this by illustrating SNR-κ scatter plots (red) and SNR-normSNR scatter plots (blue) in log-log scales. These plots suggest that the SNR is largely driven by the directional uniformity.