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 $\kappa$ to characterize the directional...
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 $x_n \Rightarrow x$ to indicate that “a random variable $x_n$ converges to $x$ in distribution.” Similarly, $x_n \xrightarrow{P} x$ means convergence in probability. An angle $\theta$ between $d$-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$ is defined by $\theta = \frac{180}{\pi} \cos^{-1} \left( \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\| \mathbf{u} \| \| \mathbf{v} \|} \right)$.

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

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

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

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

$$\hat{\mu} = \frac{\sum_{i=1}^n x_i}{\|x\|} \quad \text{and} \quad \hat{\kappa} \approx \frac{\hat{\mu}(d-1)}{d-2}$$ where $x_i$'s are random samples from the VMF distribution and $\{x_i\}$

$\|x\|$ is drawn from the VMF distribution and $\{x_i\}$.

3 THEORECTICAL MOTIVATION

3.1 ANALYSIS OF THE GRADIENT NORM STOCHASTICITY

It is a usual practice for SGD to use a minibatch gradient $\hat{g}(w) = -\nabla_w f_1(w)$ instead of a full batch gradient $g(w) = -\nabla_w f(w)$. The minibatch index set $I$ is drawn from $\{1, \ldots, n\}$ randomly. $\hat{g}(w)$ satisfies $\mathbb{E}g(w) = g(w)$ and $\text{Cov}(\hat{g}(w), g(w)) \approx \frac{1}{m} \sum_{i=1}^m g_i(w)g_i(w)^\top$ for $n \gg m$ where $n$ is the number of full data and $g_i(w) = -\nabla_w f_i(w)$ (Hoffer et al., 2017). As the batch size $m$ increases, the randomness of $\hat{g}(w)$ decreases. Hence $\mathbb{E}\|\hat{g}(w)\|$ tends to $\|g(w)\|$, and $\text{Var}(\|\hat{g}(w)\|)$, which is the variance of the norm of the minibatch gradient, vanishes. The convergence rate analysis is as the following:

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

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

and

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

Hence,

$$\sqrt{\text{Var}(\|\hat{g}(w)\|)} \leq \frac{\sqrt{2(n-m)}}{m(n-1)} \times \gamma \leq \frac{\sqrt{2(n-m)}}{m(n-1)} \times \|g(w)\|.$$ (1)

**Proof.** See Supplemental[A]

According to **Theorem 1** the large batch size $m$ reduces the variance of $\|\hat{g}(w)\|$ centered at $\mathbb{E}\|\hat{g}(w)\|$ 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))
Figure 3: (a) Asymptotic angle densities (2) of \( \theta(u, v) = \frac{180}{\pi}\cos^{-1}\langle u, v \rangle - 90 \) where \( u \) and \( v \) are independent uniformly random unit vectors for each large dimension \( d \). As \( d \to \infty \), \( \theta(u, v) \) 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 \( \theta(u, \hat{g}_j(w)) \) for 3000 minibatch gradients (black) at \( w = w_0^{(b)} \) and \( w = w_{\text{final}}^{(b)} \) with training accuracy of > 99.9\%. (c) when \( u \) is given. After SGD iterations, the density of \( \theta(u, \hat{g}_j(w)) \) converges to an asymptotic density (red). The dimension of FNN is 635,200.

Inversely proportional to \( \|g(w)\| \). This implies that the learning dynamics and convergence of SGD, measured in terms of the vanishing gradient, i.e., \( \sum_{i=1}^{n_b} \hat{g}_i(w) \approx 0 \), is not necessarily explained by the vanishing norms of minibatch gradients, but rather by the balance of the directions of \( \hat{g}_i(w) \)'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 asymptotic behavior of angles between uniformly random unit vectors in a high-dimensional space.

**Theorem 2.** Suppose that \( u \) and \( v \) are mutually independent \( d \)-dimensional uniformly random unit vectors. Then,

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

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 increasingly 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} \left( 90, \frac{(180}{\pi \sqrt{d}} \right)^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.
Moreover, \( \hat{\epsilon} \) for all \( \bar{\epsilon} \) by \( \hat{\epsilon} = \bar{\epsilon} \) for any \( \bar{\epsilon} > 0 \). If we consider \( \hat{\epsilon} = h(u) \) as a function of \( u = \| \sum_{i=1}^{n_b} x_i \| \), 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
\[
\hat{\epsilon}(w) = h \left( \frac{\sum_{i=1}^{n_b} \frac{p_i - w}{\|p_i - w\|}}{1 - \bar{\epsilon}^2} \right),
\]
which is measured from the directions from the current location \( w \) to the fixed points \( p_i \)'s, where \( h(\cdot) \) is a function defined in Lemma 1. Since \( h(\cdot) \) is an increasing function, we may focus only on \( \| \sum_{i=1}^{n_b} \frac{p_i - w}{\|p_i - w\|} \| \) to see how \( \hat{\epsilon} \) behaves with respect to its argument. Lemma 2 implies that the estimated uniformity \( \hat{\epsilon} \) increases if we move away from \( w_0 \) to \( w' = w_0 + \epsilon \sum_{i=1}^{n_b} \frac{p_i - w_0}{\|p_i - w_0\|} \) with a small \( \epsilon \) (Figure 4(a)). In other words, \( \hat{\epsilon}(w') < \hat{\epsilon}(w_0) \).

Lemma 2. Let \( p_1, p_2, \ldots, p_{n_b} \) be \( d \)-dimensional vectors. If all \( p_i \)'s are not on the ray from the current location \( w \), then there exists positive number \( \eta \) such that
\[
\left\| \frac{\sum_{j=1}^{n_b} p_j - w - \epsilon \sum_{i=1}^{n_b} \frac{p_i - w}{\|p_i - w\|}}{\|p_j - w - \epsilon \sum_{i=1}^{n_b} \frac{p_i - w}{\|p_i - w\|}\|} \right\| < \sum_{i=1}^{n_b} \frac{p_i - w}{\|p_i - w\|}
\]
for all \( \epsilon \in (0, \eta] \).

Proof. See Supplemental C.2

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 \( \kappa \) 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 \( \hat{\kappa} \) is almost parallel to the direction of full batch gradient. In other words, \( \kappa \) 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 \( \kappa \).

Lemma 1. The approximated estimator of \( \kappa \) induced from the \( d \)-dimensional unit vectors \( \{x_1, x_2, \ldots, x_{n_b}\} \),
\[
\hat{\kappa} = \frac{\bar{r}(d - \bar{r}^2)}{1 - \bar{r}^2},
\]
is a strictly increasing function on \([0, 1]\), where \( \bar{r} = \| \sum_{i=1}^{n_b} x_i \| \). If we consider \( \hat{\kappa} = h(u) \) as a function of \( u = \| \sum_{i=1}^{n_b} x_i \| \), 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.
We make the connection between the observation above and SGD by first viewing $p_i$’s as local minibatch solutions.

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

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

Each epoch of SGD with a learning rate $\eta$ computes a series of $w^i_j = w^0 + \eta \sum_{j=1}^{3} \hat{g}_i(w^i_{j-1})$ for all $j \in \{1, \ldots, n_b\}$. If $-\nabla_w f(\cdot)$ is Lipschitz continuous, the negative gradient of the $i$-th iteration in the $t$-th epoch satisfies $\hat{g}_i(w^i_{t-1}) \approx \hat{g}_i(w^0)$ for a small $\eta$. Moreover, Theorem 1 implies $\|\hat{g}_i(w^i_{t-1})\| \approx \|\hat{g}_i(w^0)\| = \tau$ for all $i \in \{1, \ldots, n_b\}$ 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 $p_i(w^0) = p_i(w^0) = p_i$ for all $i = 1, 2, 3$. Then,

$$\hat{\kappa}(w^0_i) = h\left(\left\{\sum_{i=1}^{3} p_i - w^0\right\} \|\|p_i - w^0\|\|\right),$$

and $\hat{\kappa}(w^0_i) \approx h\left(\left\{\sum_{i=1}^{3} p_j - w^0 - \eta \sum_{i=1}^{3} \hat{g}_i(w^i_{j-1})\right\} \|\|p_j - w^0 - \eta \sum_{i=1}^{3} \hat{g}_i(w^i_{j-1})\|\|\right)$. If $\hat{g}_i(w^0)$ is parallel to $p_i - w^0$ for each $i$,

$$\sum_{j=1}^{3} \frac{p_j - w^0 - \eta \sum_{i=1}^{3} \hat{g}_i(w^i_{j-1})}{\|p_j - w^0 - \eta \sum_{i=1}^{3} \hat{g}_i(w^i_{j-1})\|} = \sum_{j=1}^{3} \frac{p_j - w^0}{\|p_j - w^0\|}.$$

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

Since $\hat{g}_i(w^0)$ may not be perfectly parallel to $p_i - w^0$, we further show that a small movement of $\eta \sum_i \frac{\hat{g}_i(w^i_{j-1})}{\|\|\hat{g}_i(w^i_{j-1})\||}$ in Theorem 3 below.

**Theorem 3.** Let $p_1(w^0), p_2(w^0), \ldots, p_{n_b}(w^0)$ be $d$-dimensional vectors, and all $p_i(w^0)$’s are not on a single ray from the current location $w^0$. If

$$\left\|\sum_{i=1}^{n_b} \frac{p_i(w^0)}{\|p_i(w^0)\|} - w^0\right\| - \eta \sum_{i=1}^{n_b} \frac{\hat{g}_i(w^i_{j-1})}{\|\|\hat{g}_i(w^i_{j-1})\||} \leq \xi$$

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

$$\left\|\sum_{i=1}^{n_b} \frac{p_i(w^0)}{\|p_i(w^0)\|} - w^0 - \epsilon \sum_{i=1}^{n_b} \frac{\hat{g}_i(w^i_{j-1})}{\|\|\hat{g}_i(w^i_{j-1})\||} \right\| < \left\|\sum_{i=1}^{n_b} \frac{p_i(w^0)}{\|p_i(w^0)\|} - w^0\right\|$$

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 $\sum_i \frac{p_i(w^0) - w^0}{\|p_i(w^0) - w^0\|}$. With additional assumptions on each minibatch loss function, we have a sufficient condition for (3), summarized in Corollary 3.1.

**Corollary 3.1.** Let $p_i$ be the local minibatch solution of each $f_{i,i}$. Suppose a region $\mathcal{R}$ satisfying:

For all $w, w' \in \mathcal{R}$, $p_i(w) = p_i(w') = p_i$

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

Without the corollary above, we need to solve $p_i(w^0_i) = \arg \min_{w \in N(w^0_i)} f_i(w)$ for all $i \in \{1, \ldots, n_b\}$, where $n_s$ is the number of samples to estimate $\kappa$, in order to compute $\hat{\kappa}(w^0_i)$. Corollary 3.1 however implies that we can compute $\hat{\kappa}(w^0_i)$ by using $\hat{g}_i(w^0_i) / \|g_i(w^0_i)\|$ instead of $\frac{p_i(w^0_i) - w^0}{\|p_i(w^0_i) - w^0\|}$, 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 $w^0$ to $w^{t+1}$. 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 $\mu$, just like the average of all $n_b$ normalized minibatch gradients. Thus, we expect with a large $n'_b$,

$$\left\| \sum_{i=1}^{n'_b} \frac{p_i(w^0_i) - w^0}{\|p_i(w^0_i) - w^0\|} - \sum_{i=1}^{n'_b} \frac{\hat{g}_i(w^{i-1})}{\|\hat{g}_i(w^{i-1})\|} \right\| \leq \xi,$$

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

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) $\uparrow \| \mathbb{E}_{\omega} \nabla_w f_\omega \| / \text{tr}(\text{Cov}(\nabla_w f_\omega, \nabla_w f_\omega)) \downarrow$
- Gradient norm stochasticity (GNS) $\uparrow \mathbb{E} \| \nabla_w f_\omega \| / \sqrt{\text{Var}(\| \nabla_w f_\omega \|)} \downarrow$
- Directional Uniformity $\uparrow \kappa \downarrow$

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 $\kappa$ decreases over training regardless of the network’s depth in Figure 5 (a,b). We however also notice that $\kappa$ 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
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. Nevertheless, as shown in Figure 5 (c), $\kappa$ eventually drops below its initial point, although this trend is not monotonic and $\kappa$ fluctuates significantly over training. The addition of batch normalization (+BN) helps with the fluctuation but $\kappa$ does not monotonically decrease (Figure 5 (g)). On the other hand, we observe the monotonic decrease of $\kappa$ 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 $\kappa$ 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 $\kappa$, implying that either or both of GNS and $\kappa$ 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 $\kappa$ 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.
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 $\kappa$. We normalized each quantity by its maximum value over training for easier comparison on a single plot. In all the cases, SNR (orange) and $\kappa$ (red) are almost entirely correlated with each other, while normSNR is less correlated. (b,d,f) We further verify this by illustrating SNR-$\kappa$ 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 [Schwartz-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


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:

\[ s_i = \begin{cases} 1, & \text{if } i \in I \\ 0, & \text{if } i \notin I \end{cases} \]

Then we have

\[ \hat{g}(w) = \frac{1}{m} \sum_{i=1}^{n} g_i(w)s_i. \]

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

\[ 0 \leq \mathbb{E}\|\hat{g}(w)\|^2 - \|g(w)\|^2 \leq \frac{2(n-m)}{m(n-1)} \gamma. \quad (5) \]

where \( \gamma = \max_{i,j \in \{1,\ldots,n\}} |\langle g_i(w), g_j(w) \rangle| \).

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

\[ \mathbb{E}\|\hat{g}(w)\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{m^2} \langle g_i(w), g_j(w) \rangle \mathbb{E}[s_is_j]. \]

Since \( \mathbb{E}[s_is_j] = \frac{m}{n} \delta_{ij} + \frac{m(m-1)}{n(n-1)} (1 - \delta_{ij}) \),

\[ \mathbb{E}\|\hat{g}(w)\|^2 - \|g(w)\|^2 = \left( \frac{1}{mn} - \frac{m-1}{mn(n-1)} \right) \sum_{i=1}^{n} \langle g_i(w), g_i(w) \rangle \\
+ \left( \frac{m-1}{mn(n-1)} - \frac{1}{n^2} \right) \sum_{i=1}^{n} \sum_{j=1}^{n} \langle g_i(w), g_j(w) \rangle \\
= \left( \frac{1}{mn} - \frac{m-1}{mn(n-1)} \right) \sum_{i=1}^{n} \langle g_i(w), g_i(w) \rangle + \frac{m-n}{mn^2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle g_i(w), g_j(w) \rangle \\
\leq \left( \frac{1}{mn} - \frac{m-1}{mn(n-1)} \right) n\gamma + \frac{m-n}{mn^2(n-1)} n^2\gamma \\
= \frac{2(n-m)}{m(n-1)} \gamma \]

where \( \gamma = \max_{i,j \in \{1,\ldots,n\}} |\langle g_i(w), g_j(w) \rangle| \).

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

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

and

\[ \text{Var}(\|\hat{g}(w)\|) \leq \frac{2(n-m)}{m(n-1)} \gamma. \]

Hence,

\[ \sqrt{\frac{\text{Var}(\|\hat{g}(w)\|)}{\mathbb{E}\|\hat{g}(w)\|^2}} \leq \frac{2(n-m)}{m(n-1)} \times \frac{\gamma}{\|g(w)\|^2}. \]
B PROOFS FOR THEOREM

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 \( \{x_n\}, \{y_n\} \) be a sequence of random variables that satisfies \( x_n \xrightarrow{P} x \) and \( y_n \xrightarrow{P} \rho \) when \( n \) goes to infinity and \( \rho \) is constant. Then \( x_n y_n \xrightarrow{P} \rho \).

**Theorem B.2. (Delta method, Casella & Berger (2002))** Let \( y_n \) be a sequence of random variables that satisfies \( \sqrt{n}(y_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \). 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 \( \sqrt{n}(f(y_n) - f(\mu)) \xrightarrow{d} \mathcal{N}(0, \sigma^2[f'(\mu)]^2) \).

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

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

\[
    u = \frac{x}{\|x\|}.
\]

Suppose that two independent uniformly random unit vector \( u \) and \( v \) are generated by two independent \( d \)-dimensional standard normal vector \( x = (x_1, x_2, \cdots, x_d) \) and \( y = (y_1, y_2, \cdots, y_d) \). Denote them

\[
    u = \frac{x}{\|x\|} \quad \text{and} \quad v = \frac{y}{\|y\|}.
\]

By SLLN, we have

\[
    \frac{\|x\|}{\sqrt{d}} \to 1 \quad \text{a.s.}
\]

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

\[
    \frac{\langle x, y \rangle}{\sqrt{d}} = \sqrt{d} \left( \frac{1}{d} \sum_{i=1}^{d} x_i y_i \right) \xrightarrow{d} \mathcal{N}(0, 1).
\]

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

\[
    \sqrt{d} \langle u, v \rangle \xrightarrow{d} \mathcal{N}(0, 1).
\]

\[\square\]
Theorem 2. Suppose that \( \mathbf{u} \) and \( \mathbf{v} \) are mutually independent \( d \)-dimensional uniformly random unit vectors. Then,

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

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

\( \square \)

C Proofs for Theorem 3

C.1 Proof of Lemma 1

Lemma 1. The approximated estimator of \( \kappa \) induced from the \( d \)-dimensional unit vectors \( \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{n_b}\} \),

\[
\hat{\kappa} = \frac{\bar{r}(d - \bar{r}^2)}{1 - \bar{r}^2},
\]
where \( \bar{r} = \frac{\| \sum_{i=1}^{n_b} \mathbf{x}_i \|}{n_b} \) is a strict increasing function on \([0, 1]\). We can also consider \( \hat{\kappa} = h(\mathbf{u}) \) as function of \( u = \| \sum_{i=1}^{n_b} \mathbf{x}_i \| \). 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 \( \| \sum_{i=1}^{n_b} \mathbf{x}_i \| \leq \sum_{i=1}^{n_b} ||\mathbf{x}_i|| = n_b \). 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,

\[
h''(u) = \frac{1}{n_b^2} \frac{d^2\hat{\kappa}}{d\bar{r}^2}
\]
\[
= \frac{2\bar{r}^5 + (4 - 8d)\bar{r}^3 + (8d - 6)\bar{r}}{n_b^2(1 - \bar{r}^2)^4} > 0
\]
due to \( \bar{r} \in [0, 1] \). Therefore \( h'(\cdot) \) is also increasing on \([0, n_b]\). \( \square \)
C.2 Proof of Lemma 2

Lemma 2. Let \( p_1, p_2, \ldots, p_n \) be d-dimensional vectors. If all \( p_i \)'s are not on the ray from the current location \( w \), then there exists positive number \( \eta \) such that

\[
\left\| \sum_{j=1}^{n_b} p_j - w - \epsilon \sum_{i=1}^{n_b} \frac{p_i - w}{\| p_i - w \|} \right\| < \left\| \sum_{i=1}^{n_b} p_i - w \right\|
\]

for all \( \epsilon \in (0, \eta) \).

Proof. Without loss of generality, we regard \( w \) as the origin. Let \( f(\epsilon) = \left\| \sum_{j=1}^{n_b} p_j - \epsilon \sum_{i=1}^{n_b} \frac{p_i - w}{\| p_i - w \|} \right\| \), then \( f(0) = \left\| \sum_{i=1}^{n_b} p_i \right\| \). Therefore, we only need to show \( f'(0) < 0 \). Now denote \( x_j = \frac{p_j}{\| p_j \|}, p_j(\epsilon) = p_j - \epsilon \sum_{i=1}^{n_b} x_i \) and \( u = -\sum_{i=1}^{n_b} x_i \). That is, \( p_j(\epsilon) = p_j + \epsilon u \). Since

\[
f(\epsilon) = \left\{ \sum_{j=1}^{n_b} \frac{p_j(\epsilon)}{\| p_j(\epsilon) \|}, \sum_{j=1}^{n_b} \frac{p_j(\epsilon)}{\| p_j(\epsilon) \|} \right\},
\]

we have

\[
f'(\epsilon) = 2 \left( \sum_{j=1}^{n_b} \frac{p_j(\epsilon)}{\| p_j(\epsilon) \|} \right) \frac{d}{d\epsilon} \left( \sum_{j=1}^{n_b} \frac{p_j(\epsilon)}{\| p_j(\epsilon) \|} \right)
\]

and

\[
\frac{d}{d\epsilon} \left( \sum_{j=1}^{n_b} \frac{p_j(\epsilon)}{\| p_j(\epsilon) \|} \right) = \sum_{j=1}^{n_b} \frac{\| p_j(\epsilon) \| u - \left( \frac{u \cdot p_j(\epsilon)}{\| p_j(\epsilon) \|} \right) p_j(\epsilon)}{\| p_j(\epsilon) \|^2}.
\]

Hence

\[
f'(\epsilon) = 2 \left( \sum_{j=1}^{n_b} \frac{p_j(\epsilon)}{\| p_j(\epsilon) \|} \right) \sum_{j=1}^{n_b} \frac{\| p_j(\epsilon) \| u - \left( \frac{u \cdot p_j(\epsilon)}{\| p_j(\epsilon) \|} \right) p_j(\epsilon)}{\| p_j(\epsilon) \|^2}.
\]

Note that \( p_j(0) = p_j \) and \( \| x_j \| = 1 \). We have

\[
f'(0) = 2 \left( \sum_{j=1}^{n_b} \frac{p_j}{\| p_j \|} \right) \sum_{j=1}^{n_b} \frac{\| p_j \| u - \left( \frac{u \cdot p_j}{\| p_j \|} \right) p_j}{\| p_j \|^2}
\]

\[
= 2 \left( \sum_{j=1}^{n_b} \frac{p_j}{\| p_j \|} \right) \sum_{j=1}^{n_b} \frac{1}{\| p_j \|} \left( u - \left( \frac{u \cdot p_j}{\| p_j \|} \right) p_j \right)
\]

\[
= 2 \left( \sum_{j=1}^{n_b} x_j, \sum_{j=1}^{n_b} \frac{1}{\| p_j \|} \left( u - \left( \frac{u \cdot x_j}{\| x_j \|} \right) x_j \right) \right)
\]

\[
= 2 \left( -u, \sum_{j=1}^{n_b} \frac{1}{\| p_j \|} \left( u - \left( \frac{u \cdot x_j}{\| x_j \|} \right) x_j \right) \right)
\]

\[
= -2 \sum_{j=1}^{n_b} \frac{\| u \|^2 - \left( \frac{u \cdot x_j}{\| x_j \|} \right)^2}{\| p_j \|}
\]

\[
\leq -2 \sum_{j=1}^{n_b} \frac{\| u \|^2 - \| u \|^2 \| x_j \|^2}{\| p_j \|}
\]

\[
= 0
\]

Since the equality holds when \( \langle u, x_j \rangle^2 = \| u \|^2 \| x_j \|^2 \) for all \( j \), we can have strict inequality when all \( p_i \)'s are not on the same ray from the origin. \( \square \)
C.3 Proof of Theorem 3

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

**Theorem 3.** Let \(p_1(w^0), p_2(w^0), \ldots, p_{n_b}(w^0)\) be \(d\)-dimensional vectors, and all \(p_i(w^0)\)'s are not on the ray from the current location \(w^0\). If

\[
\left\| \sum_{i=1}^{n_b} \frac{p_i(w^0_i) - w^0_i}{\|p_i(w^0) - w^0\|} - \sum_{i=1}^{n_b} \frac{\tilde{g}_i(w^{i-1})}{\|\tilde{g}_i(w^{i-1})\|} \right\| \leq \xi
\]

(6)

for sufficiently small \(\xi > 0\), then there exists positive number \(\eta\) such that

\[
\left\| \sum_{j=1}^{n_b} \frac{p_j(w^0_j) - w^0_j}{\|p_j(w^0) - w^0\|} - \epsilon \sum_{j=1}^{n_b} \frac{\tilde{g}_j(w^{j-1})}{\|\tilde{g}_j(w^{j-1})\|} \right\| < \left\| \sum_{i=1}^{n_b} \frac{p_i(w^0_i) - w^0_i}{\|p_i(w^0) - w^0\|} \right\|
\]

(7)

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

**Proof.** Similarly, we regard \(w^0\) as the origin \(0\). For simplicity, write \(p_i(0)\) and \(\tilde{g}_i(w^{i-1})\) as \(p_i\) and \(\tilde{g}_i\), respectively. Let \(f(\epsilon) = \left\| \sum_{j=1}^{n_b} \frac{p_j - \epsilon \frac{p_j}{\|p_j\|} \frac{\tilde{g}_j(w^{j-1})}{\|\tilde{g}_j(w^{j-1})\|}}{\|p_j\|} \right\|^2\) and \(\tilde{f}(\epsilon) = \left\| \sum_{j=1}^{n_b} \frac{p_j - \epsilon \frac{p_j}{\|p_j\|} \frac{\tilde{g}_j(w^{j-1})}{\|\tilde{g}_j(w^{j-1})\|}}{\|p_j\|} \right\|^2\).

Denote \(u = -\sum_{j=1}^{n_b} \frac{p_j}{\|p_j\|}, t = \sum_{i=1}^{n_b} \frac{p_i}{\|p_i\|} - \sum_{i=1}^{n_b} \frac{\tilde{g}_i}{\|\tilde{g}_i\|}\) and \(\tilde{p}_j(\epsilon) = p_j + \epsilon(u + t)\). Then

\[
\tilde{f}(\epsilon) = \left\| \sum_{j=1}^{n_b} \frac{\tilde{p}_j(\epsilon)}{\|\tilde{p}_j(\epsilon)\|} \right\|^2.
\]

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

\[
f''(\epsilon) = 2 \left\langle \sum_{j=1}^{n_b} \frac{\tilde{p}_j(\epsilon)}{\|\tilde{p}_j(\epsilon)\|}, \sum_{j=1}^{n_b} \frac{\tilde{p}_j(\epsilon) \left( (u + t) - \frac{(u + t, p_j)}{\|p_j\|^2} p_j \right)}{\|p_j\|^2} \right\rangle.
\]

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

\[
f''(0) = 2 \left\langle \sum_{j=1}^{n_b} \frac{p_j}{\|p_j\|}, \sum_{j=1}^{n_b} \frac{p_j \left( (u + t) - \frac{(u + t, p_j)}{\|p_j\|^2} p_j \right)}{\|p_j\|^2} \right\rangle
\]

\[
= 2 \left\langle -u, \sum_{j=1}^{n_b} \frac{1}{\|p_j\|} \left( u + t - \langle u + t, x_j \rangle x_j \right) \right\rangle
\]

\[
= 2 \left\langle -u, \sum_{j=1}^{n_b} \frac{1}{\|p_j\|} \left( u - \langle u, x_j \rangle x_j \right) \right\rangle + 2 \left\langle -u, \sum_{j=1}^{n_b} \frac{1}{\|p_j\|} \left( t - \langle t, x_j \rangle x_j \right) \right\rangle
\]

\[
f''(0) = 2 \left\langle (u, t) - \langle t, x_j \rangle \langle u, x_j \rangle \right\rangle
\]

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

\[
f''(0) < 0 \iff 2 \sum_{j=1}^{n_b} \frac{1}{\|p_j\|} \left( \langle t, x_j \rangle \langle u, x_j \rangle - \langle u, t \rangle \right) < |f'(0)|.
\]
3.1 Recall that $R_L$ specifies the direction of the minibatch gradient with the corresponding local minibatch solution. Since both the batch size is sufficiently large and the learning rate for all $\Corollary 3.1$. Let $\|w\| = \min_j \|p_j\|$. If $\xi_\epsilon < \frac{f'(0)\epsilon}{2n_b}$, then $\tilde{f}(\epsilon) < 0$.

D PROOFS FOR COROLLARY 3.1

\Corollary 3.1 Let $p_i$'s be local minibatch solutions for each $f_i$'s. Suppose that a region $\mathcal{R}$ satisfies:

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

Proof. Recall that $w^0_{t+1} = w^0_t + \eta \sum_{j=1}^{n_b} g_j(w_{t}^{i-1})$ where $\eta$ is a learning rate. To prove Corollary 3.1 we need to show $\tilde{k}(w^{t+1}_t) < \tilde{k}(w^0_t)$ which is equivalent to

$$\left\| \sum_{j=1}^{n_b} p_{j}(w^0_{t+1}) - w^0_t - \eta \sum_{j=1}^{n_b} g_j(w^0_{t+1}) \right\| < \left\| \sum_{i=1}^{n_b} p_i(w^0_t) - w^0_t \right\|.$$  \((8)\)

Since $\|\nabla w^2 f_i(\cdot)\|_2$ is bounded on $\mathcal{R}$, $\nabla w f_i(\cdot)$ is Lipschitz continuous on $\mathcal{R}$ (Bottou, 2010). If the batch size is sufficiently large and the learning rate $\eta$ is sufficiently small, $\|g_i(w^0_t)\| \approx \tau$ for all $i$. Therefore, we have

$$\eta \sum_{i=1}^{n_b} g_i(w^0_{t+1}) \approx \tau \eta \sum_{i=1}^{n_b} \|g_i(w^0_{t+1})\|$$

If we denote $\tau \eta$ as $\epsilon$, we can convert $\text{(5)}$ to $\text{(2)}$.

$$\left\| \sum_{j=1}^{n_b} p_{j}(w^0_{t+1}) - w^0_t - \epsilon \sum_{i=1}^{n_b} g_i(w^0_{t+1}) \right\| < \left\| \sum_{i=1}^{n_b} p_i(w^0_t) - w^0_t \right\|.$$  \((9)\)

Since both $w^0_{t+1}$ and $w^0_t$ are in $\mathcal{R}$ for small learning rate, we have $p_i(w^0_{t+1}) = p_i(w^0_t) = p_i$ by the assumption. That is, $\text{(9)}$ is equivalent to $\text{(7)}$. In $\text{(7)}$, $\|g_i(w^0_{t+1})\|$ cannot be replaced by $\|p_i(w^0_t) - p_i\|$. 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 $c(A)$ of a matrix $A$ is defined as
\[ c(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \]
where $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ are maximal and minimal singular values of $A$, respectively. If $A$ is positive-definite matrix, then
\[ c(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}. \]
Here $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are maximal and minimal eigenvalues of $A$, respectively.

Lemma D.1. If the condition number of the positive definite Hessian matrix of $f_{\hat{w}}$ at the local minibatch solution, $p_i$, denoted by $H_i = \nabla_w^2 f_\hat{w}(p_i)$ is approximately $1$(well-conditioned), then the direction to $p_i$ from $w$ is almost parallel to its negative gradient at $w$. That is, for all $w \in \mathbb{R}$,
\[ \left\| \frac{p_i - w}{\|p_i - w\|} - \frac{\hat{g}_i(w)}{\|\hat{g}_i(w)\|} \right\| \approx 0 \]
where $\hat{g}_i(w) = -\nabla_w f_{\hat{w}}(w)$.

Proof. By the second order Taylor expansion,
\[ f_\hat{w}(w) \approx f_\hat{w}(p_i) + \frac{1}{2}(w - p_i)^\top H_i(w - p_i). \]
Hence,
\[ \hat{g}_i(w) = -\nabla_w f_{\hat{w}}(w) \approx -H_i(w - p_i) \]
Denote $p_i - w$ as $x$. Then, we only need to show
\[ \left\| \frac{x}{\|x\|} - \frac{H_i x}{\|H_i x\|} \right\|^2 \approx 0 \]
Note that $H_i$ is positive definite, so we can diagonalize it as $H_i = P_i^\top \Lambda_i P_i$ where $P_i$ is an orthonormal transition matrix for $H_i$.
\[
\left\| \frac{x}{\|x\|} - \frac{H_i x}{\|H_i x\|} \right\|^2 = 2 - 2 \frac{x^\top H_i x}{\|x\| \cdot \|H_i x\|} \\
= 2 - 2 \frac{(P_i x)^\top \Lambda_i P_i x}{\|P_i x\| \cdot \|P_i \Lambda_i P_i x\|} \\
= 2 - 2 \frac{(P_i x)^\top \Lambda_i P_i x}{\|P_i x\| \cdot \|P_i \Lambda_i P_i x\|} \\
\leq 2 - 2 \frac{\sum_j \lambda_j (P_i x)_j^2}{\|P_i x\| \cdot \|P_i \Lambda_i P_i x\|} \\
\leq 2 - 2 \frac{\lambda_{\min} \|P_i x\|^2}{\|P_i \| \cdot \|P_i \Lambda_i P_i x\|} \\
= 2 - 2 \frac{\lambda_{\min}}{\lambda_{\max}} \approx 0 \]
\[
\square
\]

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

Lemma D.2. Let $w$ be a parameter in $\mathbb{R}$. If the condition number of Hessian matrix of $f_{\hat{w}}$ is sufficiently near $1$(well-conditioned) and $\|w - w^0\|$ is sufficiently near $0$, then
\[
\left\| \frac{p_i - w^0}{\|p_i - w^0\|} - \frac{\hat{g}_i(w)}{\|\hat{g}_i(w)\|} \right\| \leq \frac{\xi}{n_b} \]
for sufficiently small $\xi$. 

Under review as a conference paper at ICLR 2019
Proof. We have
\[
\| p_i - w_i^0 \| \leq \frac{\hat{g}_i(w) - \hat{g}_i(w)}{\|g(w)\|} + \frac{\|p_i - w\|}{\|\hat{g}_i(w)\|} + \frac{\|p_i - w\|}{\|\hat{g}_i(w)\|} = \frac{\|p_i - w\|}{\|\hat{g}_i(w)\|},
\]
and
\[
\| p_i - w_i^0 \| \leq \frac{\|p_i - w\|}{\|\hat{g}_i(w)\|} = \frac{\|p_i - w\|}{\|\hat{g}_i(w)\|},
\]
for sufficiently small \( \epsilon \).

(See Lemma D.1.) Now we only need to show
\[
\sqrt{2 \left( 1 - \frac{\|p_i - w_i^0 \|^2 - \langle p_i - w_i^0, w - w_i^0 \rangle}{\|p_i - w_i^0\| \|p_i - w\|} \right)} < \epsilon.
\]
Since \( \|w - w_i^0\| \) is sufficiently small, we have
\[
\|p_i - w_i^0\|^2 - \langle p_i - w_i^0, w - w_i^0 \rangle = \|p_i - w_i^0\| \left( \|p_i - w_i^0\| - \langle p_i - w_i^0, w - w_i^0 \rangle \right)
\geq \|p_i - w_i^0\| \left( \|p_i - w_i^0\| - \|w - w_i^0\| \right)
= \|p_i - w_i^0\| \left( 1 - \frac{\|w - w_i^0\|}{\|p_i - w_i^0\|} \right)
\geq 0.
\]
By using the above non-negativeness, we have the following inequality.
\[
1 \geq \frac{\|p_i - w_i^0\|^2 - \langle p_i - w_i^0, w - w_i^0 \rangle}{\|p_i - w_i^0\| \|p_i - w\|}
\geq \frac{\|p_i - w_i^0\|^2 - \langle p_i - w_i^0, w - w_i^0 \rangle}{\|p_i - w_i^0\| \left( \|p_i - w_i^0\| + \|w - w_i^0\| \right)}
= \frac{\|p_i - w_i^0\| - \langle p_i - w_i^0, w - w_i^0 \rangle}{1 + \frac{\|p_i - w_i^0\|}{\|w - w_i^0\|}}
\geq \frac{\|p_i - w_i^0\|}{\|w - w_i^0\|} - 1
\]
As \( \frac{\|w - w_i^0\|}{\|p_i - w_i^0\|} \to 0^+ \), (11) is monotonically increasing to 1. This implies that (10) holds for sufficiently small \( \epsilon \).

With small learning rate, \( w_i^{i-1}'s \) are in \( R \) for all \( i \in \{1, \ldots, n_b\} \). As a result, by Lemma D.2 we have
\[
\| p_i - w_i^0 \| \leq \frac{\|p_i - w\|}{\|\hat{g}_i(\{w_i^{i-1}\})\|} \leq \frac{\epsilon}{n_b},
\]
for sufficiently small \( \epsilon \). This implies (6) since
\[
\| \sum_{i=1}^{n_b} p_i \| \leq \sum_{i=1}^{n_b} \frac{\hat{g}_i(\{w_i^{i-1}\})}{\|\hat{g}_i(\{w_i^{i-1}\})\|} \leq \sum_{i=1}^{n_b} \frac{p_i - w}{\|\hat{g}_i(\{w_i^{i-1}\})\|}.\]
Then we can apply Theorem 3 and \( \hat{\kappa}(w_{i}^{0}) > \hat{\kappa}(w_{i+1}^{0}) \) holds.

For the last statement, "Moreover, we can estimate \( \hat{\kappa}(w_{i}^{0}) \) and \( \hat{\kappa}(w_{i+1}^{0}) \) by minibatch gradients on \( w_{i}^{0} \) and \( w_{i+1}^{0} \), respectively.", recall that

\[
\hat{\kappa}(w_{i}^{0}) = h\left(\left\| \sum_{i=1}^{n_{b}} p_{i}(w_{0}^{i}) - w_{i}^{0} \right\| \right)
\]

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

\[
\left\| \frac{p_{i}(w_{0}^{i}) - w_{i}^{0}}{p_{i}(w_{0}^{i}) - w_{0}^{i}} \right\| - \frac{\hat{g}_{i}(w_{0}^{i})}{\|\hat{g}_{i}(w_{0}^{i})\|} \right\| < \frac{\xi}{n_{b}}
\]

for sufficiently small \( \xi > 0 \). Therefore,

\[
\left\| \sum_{i=1}^{n_{b}} \frac{p_{i}(w_{0}^{i}) - w_{i}^{0}}{p_{i}(w_{0}^{i}) - w_{0}^{i}} \right\| - \left\| \sum_{i=1}^{n_{b}} \frac{\hat{g}_{i}(w_{0}^{i})}{\|\hat{g}_{i}(w_{0}^{i})\|} \right\| \leq \left\| \sum_{i=1}^{n_{b}} \frac{p_{i}(w_{0}^{i}) - w_{i}^{0}}{p_{i}(w_{0}^{i}) - w_{0}^{i}} - \sum_{i=1}^{n_{b}} \frac{\hat{g}_{i}(w_{0}^{i})}{\|\hat{g}_{i}(w_{0}^{i})\|} \right\|
\]

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

\[
| h\left(\left\| \sum_{i=1}^{n_{b}} \frac{p_{i}(w_{0}^{i}) - w_{i}^{0}}{p_{i}(w_{0}^{i}) - w_{0}^{i}} \right\| \right) - h\left(\left\| \sum_{i=1}^{n_{b}} \frac{\hat{g}_{i}(w_{0}^{i})}{\|\hat{g}_{i}(w_{0}^{i})\|} \right\| \right) | \to 0
\]

as \( \xi \to 0 \). That is,

\[
\hat{\kappa}(w_{i}^{0}) \approx h\left(\left\| \sum_{i=1}^{n_{b}} \frac{\hat{g}_{i}(w_{0}^{i})}{\|\hat{g}_{i}(w_{0}^{i})\|} \right\| \right).
\]

Since \( t \) is arbitrary, we can apply this for all other \( w \in \mathcal{R} \) including \( w_{i+1}^{0} \). \( \square \)

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 \( 3 \times 3 \) convolution layer and the number of output filters is 16. After that, we stack of \( \{4, 4, 4\} \) layers with \( 3 \times 3 \) 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-connected 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 \( \kappa \) are strongly correlated while GNS (inverse of normSNR) is less correlated to GS.

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 \( \kappa \). We normalized each quantity by its maximum value over training for easier comparison on a single plot. In all the cases, SNR (orange) and \( \kappa \) (red) are almost entirely correlated with each other, while normSNR is less correlated. (b,d,f) We further verify this by illustrating SNR-\( \kappa \) 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.
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 $\kappa$. We normalized each quantity by its maximum value over training for easier comparison on a single plot. In all the cases, SNR (orange) and $\kappa$ (red) are almost entirely correlated with each other, while normSNR is less correlated. (b,d,f) We further verify this by illustrating SNR-$\kappa$ 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.
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 \( \kappa \). We normalized each quantity by its maximum value over training for easier comparison on a single plot. In all the cases, SNR (orange) and \( \kappa \) (red) are almost entirely correlated with each other, while normSNR is less correlated. (b,d,f) We further verify this by illustrating SNR-\( \kappa \) 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.
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 $\kappa$. We normalized each quantity by its maximum value over training for easier comparison on a single plot. In all the cases, SNR (orange) and $\kappa$ (red) are almost entirely correlated with each other, while normSNR is less correlated. (b,d,f) We further verify this by illustrating SNR-$\kappa$ 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.