Understanding Gradient Clipping in Private SGD:
A Geometric Perspective

Xiangyi Chen∗ Zhiwei Steven Wu† Mingyi Hong‡
June 20, 2020

Abstract

Deep learning models are increasingly popular in many machine learning applications where the training data may contain sensitive information. To provide formal and rigorous privacy guarantee, many learning systems now incorporate differential privacy by training their models with (differentially) private SGD. A key step in each private SGD update is gradient clipping that shrinks the gradient of an individual example whenever its ℓ2 norm exceeds some threshold. We first demonstrate how gradient clipping can prevent SGD from converging to stationary point. We then provide a theoretical analysis that fully quantifies the clipping bias on convergence with a disparity measure between the gradient distribution and a geometrically symmetric distribution. Our empirical evaluation further suggests that the gradient distributions along the trajectory of private SGD indeed exhibit symmetric structure that favors convergence. Together, our results provide an explanation why private SGD with gradient clipping remains effective in practice despite its potential clipping bias. Finally, we develop a new perturbation-based technique that can provably correct the clipping bias even for instances with highly asymmetric gradient distributions.

1 Introduction

Many modern applications of machine learning rely on datasets that may contain sensitive personal information, including medical records, browsing history, and geographic locations. To protect the private information of individual citizens, many machine learning systems now train their models subject to the constraint of differential privacy [Dwork et al., 2006], which informally requires that no individual training example has a significant influence on the trained model. To achieve this formal privacy guarantee, one of the most popular training methods, especially for deep learning, is differentially private stochastic gradient descent (DP-SGD) [Bassily et al., 2014, Abadi et al., 2016b, Song et al., 2013]. At a high level, DP-SGD is a simple modification of SGD that makes each step differentially private with the Gaussian mechanism: at each iteration t, it first computes a gradient estimate \( g_t \) based on a random subsample, and then updates the model using a noisy gradient \( \tilde{g}_t = g_t + \eta \), where \( \eta \) is a noise vector drawn from a multivariate Gaussian distribution.

∗Electrical and Computer Engineering Department, University of Minnesota. email: chen5719@umn.edu
†Computer Science and Engineering Department, University of Minnesota. email: zsw@umn.edu
‡Electrical and Computer Engineering Department, University of Minnesota. email: mhong@umn.edu
Despite the simple form of DP-SGD, there is a major disparity between its theoretical analysis and practical implementation. The formal privacy guarantee of Gaussian mechanism requires that the per-coordinate standard deviation of the noise vector \( \eta \) scales linearly with the \( \ell_2 \) sensitivity of the gradient estimate \( g_t \)—that is, the maximal change on \( g_t \) in \( \ell_2 \) distance if by changing a single example. To bound the \( \ell_2 \)-sensitivity, existing theoretical analyses typically assume that the loss function is \( L \)-Lipschitz in the model parameters, and the constant \( L \) is known to the algorithm designer for setting the noise rate [Bassily et al., 2014, Wang and Xu, 2019]. Since this assumption implies that the gradient of each example has \( \ell_2 \) norm bounded by \( L \), any gradient estimate from averaging over the gradients of \( m \) examples has \( \ell_2 \)-sensitivity bounded by \( L/m \). However, in many practical settings, especially those with deep learning models, such Lipschitz constant or gradient bounds are not a-priori known or even computable (since it involves taking the worst case over both examples and pairs of parameters). In practice, the bounded \( \ell_2 \)-sensitivity is ensured by gradient clipping [Abadi et al., 2016b] that shrinks an individual gradient whenever its \( \ell_2 \) norm exceeds certain threshold \( c \). More formally, given any gradient \( g \) on a simple example and a clipping threshold \( c \), the gradient clipping does the following

\[
\text{clip}(g, c) = g \cdot \max \left( 1, \frac{c}{\|g\|} \right).
\]

However, the clipping operation can create a substantial bias in the update direction. To illustrate this clipping bias, consider the following two optimization problems even without the privacy constraint.

**Example 1.** Consider optimizing \( f(x) = \frac{1}{3} \sum_{i=1}^{3} \left( x - a_i \right)^2 \) over \( x \in \mathbb{R} \), where \( a_1 = a_2 = -3 \) and \( a_3 = 9 \). Since the gradient \( \nabla f(x) = x - 1 \), the optimum is \( x^* = 1 \). Now suppose we run SGD with gradient clipping with a threshold of \( c = 1 \). At the optimum, the gradients for all three examples are clipped and the expected clipped gradient is \( 1/3 \), which leads the parameter to move away from \( x^* \).

**Example 2.** Let \( f(x) = \frac{1}{2} \sum_{i=1}^{2} \left( x - a_i \right)^2 \), where \( a_1 = -3 \) and \( a_2 = 3 \). The minimum of \( f \) is achieved at \( x^* = 0 \), where the expected clipped gradient is also 0. However, SGD with clipped gradients and \( c = 1 \) may never converge to \( x^* \) since the expected clipped gradients are all 0 for any \( x \in [-2, 2] \), which means all these points are "stationary" for the algorithm.

Both examples above show that clipping bias can prevent convergence in the worst case. Existing analyses on gradient clipping quantify this clipping bias either with 1) the difference between clipped and unclipped gradients [Pichapati et al., 2019], or 2) the fraction of examples with gradient norms exceeding the clip threshold \( c \) [Zhang et al., 2019]. These approaches suggest that a small clip threshold will lead to large clipping bias and worsen the training performance of DP-SGD. However, in practice, DP-SGD often remains effective even with a small clip threshold [Beaulieu-Jones et al., 2019, Bu et al., 2019], which indicates a gap in the current theoretical understanding of gradient clipping.

### 1.1 Our results

We study the effects of gradient clipping on SGD and DP-SGD and provide:

**Symmetry-based analysis.** We characterize the clipping bias on the convergence to stationary points through the geometric structure of the gradient distribution. To isolate the clipping
effects, we first analyze the non-private SGD with gradient clipping (but without Gaussian perturbation), with the following key analysis steps. 1) We first show that the inner product $\mathbb{E}[(\nabla f(x_t), g_t)]$ goes to zero in SGD, where $\nabla f(x)$ denotes the true gradient and $g_t$ denotes a clipped stochastic gradient. 2) We then show that when the gradient distribution is symmetric, inner product upper bounds a constant re-scaling of $\|\nabla f(x_t)\|$, and so SGD minimizes the gradient norm. 3) We quantify the clipping bias via a coupling between the gradient distribution and a nearby symmetric distribution and express it as a disparity measure (that resembles the Wasserstein distance) between the two distributions. As a result, when the gradient distributions are near-symmetric or when the clipping bias favors convergence, the clipped gradient remains aligned with the true gradient, even if clipping aggressively shrinks almost all the sample gradients.

**Theoretical and empirical evaluation of DP-SGD.** Building on the previous SGD analysis, we obtain a similar convergence guarantee on DP-SGD with gradient clipping. Importantly, we are able to prove such convergence guarantee even without Lipschitzness of the loss function, which is often required for DP-SGD analyses. We also provide extensive empirical studies to investigate the gradient distributions of DP-SGD across different epochs on two real datasets. To visualize the symmetricity of the gradient distributions, we perform multiple random projections on the gradients and examine the two-dimensional projected distributions. Our results suggest that the gradient distributions in DP-SGD quickly exhibit symmetricity, despite the asymmetry at initialization.

**Gradient correction mechanism.** Finally, we provide a simple modification to DP-SGD that can mitigate the clipping bias. We show that perturbing the gradients before clipping can provably reduce the clipping bias for any gradient distribution. The pre-clipping perturbation does not by itself provide privacy guarantees, but can trade-off the clipping bias with higher variance.

### 1.2 Related work

The divergence caused by the clipping bias was also studied by prior work. In Pichapati et al. [2019], an adaptive gradient clipping method is analyzed and the divergence is characterized by a bias depending on the difference between the clipped and unclipped gradients. However, they study a different variant of clipping that bounds the $\ell_\infty$ norm of the gradient instead of $\ell_2$ norm; the latter, which we study in this paper, is the more commonly used clipping operation [Abadi et al., 2016b,a]. In Zhang et al. [2019], the divergence is characterized by a bias depending on the clipping probability. These results suggest that, the clipping probability as well as the bias are inversely proportional to the size of the clipping threshold. For example, small clipping threshold results in large bias in the gradient estimation, which can potentially lead to worse training and generalization performance. Thakkar et al. [2019] provides another adaptive gradient clipping heuristic that sets the threshold based on a privately estimated quantile, which can be viewed as minimizing the clipping probability.

### 2 Convergence of SGD with clipped gradient

In this section, we analyze convergence of SGD with clipped gradient, but without the Gaussian perturbation. This simplification is useful for isolating the clipping bias. Consider the standard
stochastic optimization formulation

\[
\min_{x \in \mathbb{R}^d} f(x) = \mathbb{E}_{s \sim D}[f(x, s)],
\]

where \(x \in \mathbb{R}^d\) is the optimization variable; \(D\) denotes the underlying distribution over the examples \(s\). In the next section, we will instantiate \(D\) as the empirical distribution over the private dataset. We assume that the algorithm is given access to a stochastic gradient oracle: given any iterate \(x_t\) of SGD, the oracle returns \(\nabla f(x_t) + \xi_t\), where \(\xi_t\) is independent noise with zero mean. In addition, we assume \(f(x)\) is \(G\)-smooth, i.e. \(\|\nabla f(x) - \nabla f(y)\| \leq G\|x - y\|, \forall x, y\). At each iteration \(t\), SGD with gradient clipping performs the following update:

\[
x_{t+1} = x_t - \alpha \text{clip}(\nabla f(x_t) + \xi_t, c) := x_t - \alpha g_t,
\]

where \(g_t := \text{clip}(\nabla f(x_t) + \xi_t, c)\) denotes the realized clipped gradient.

To carry out the analysis of iteration \(3\), we first note that the standard convergence analysis for SGD-type method consists of two main steps:

- **S1** Show that the term \(\mathbb{E}[\langle \nabla f(x_t), g_t \rangle]\) diminishes to zero.
- **S2** Show that the aforementioned quantity is proportional to \(\|\nabla f(x_t)\|^2\) or \(c\|\nabla f(x_t)\|\), indicating that the size of gradient also decreases to zero.

In our analysis below, we will see that showing the first step is relatively easy, while the main challenge is to show that the second step holds true. Our first result is given below.

**Theorem 1.** Let \(G\) be the Lipschitz constant of \(\nabla f\) such that \(\|\nabla f(x) - \nabla f(y)\| \leq G\|x - y\|, \forall x, y\). For SGD with gradient clipping of threshold \(c\), if we set \(\alpha = \frac{1}{\sqrt{T}}\), we have

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\langle \nabla f(x_t), g_t \rangle] \leq \frac{D_f}{\sqrt{T}} + \frac{G}{2\sqrt{T}}c^2,
\]

where \(D_f := f(x_1) - \min_x f(x)\).

Note that for SGD without clipping, we have \(\mathbb{E}[\langle \nabla f(x_t), g_t \rangle] = \|\nabla f(x_t)\|^2\), so the convergence can be readily established. However, when clipping is applied, the expectation is different but if we have \(\mathbb{E}[\langle \nabla f(x_t), g_t \rangle]\) being positive, or have it to scale with \(\|\nabla f(x_t)\|\), we can still establish a convergence guarantee. However, the divergence examples (Example 1 and 2) indicate proving this second step requires additional conditions. Now we study a geometric condition that is observed empirically.

### 2.1 Symetricity-Based Analysis on Gradient Distribution

Let \(p_t(\xi_t)\) be the probability density function of \(\xi_t\) and \(\hat{p}_t(\xi_t)\) is an arbitrary distribution. To quantify the clipping bias, we start the analysis with the following decomposition:

\[
\mathbb{E}_{\xi_t \sim p_t}[\langle \nabla f(x_t), g_t \rangle] = \mathbb{E}_{\xi_t \sim \hat{p}_t}[\langle \nabla f(x_t), g_t \rangle] + \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle (p_t(\xi_t) - \hat{p}_t(\xi_t))d\xi_t, \quad \text{:=} b_t
\]
In (5), we can choose \( \tilde{p}(\xi_t) \) to be some "nice" distribution that can effectively relate \( \mathbb{E}_{\xi_t \sim \tilde{p}}[\langle \nabla f(x_t), g_t \rangle] \) to \( \|\nabla f(x_t)\|^2 \) and the remaining term will be treated as the bias. This way of splitting ensures that when the gradients follow a "nice" distribution, the bias will diminish with the distance between \( p \) and \( \tilde{p}_t \). More precisely, we want to find a distribution \( \tilde{p} \) such that \( \mathbb{E}_{\xi_t \sim \tilde{p}}[\langle \nabla f(x_t), g_t \rangle] \) is lower bounded by norm squared of the true gradient and thus convergence can be ensured.

A straightforward "nice" distribution will be \( \langle \nabla f(x_t), g_t \rangle \geq \Omega(\|\nabla f(x_t)\|^{\frac{3}{2}}) \), \( \forall g_t \), i.e. all stochastic gradients are positively aligned with the true gradient. This may be satisfied when the gradient is large and the noise \( \xi \) is bounded. However, when the gradient is small, it is hard to argue that this can still be true in general. Specifically, in the training of neural nets, the cosine similarities between many stochastic gradients and the true gradient (i.e. \( \cos(\nabla f(x_t), \nabla f(x_t) + \xi_t) \)) can be negative, which implies that this assumption does not hold (see Figure 3 in Section 4).

Although Figure 3 seems to exclude the ideal distribution, we observe that the distribution of cosine of the gradients appears to be symmetric. Will such a "symmetry" property help define the "nice" distribution for gradient clipping? If so, how to characterize the performance of gradient clipping in this situation? In the following result, we rigorously answer to these questions.

**Theorem 2.** Assume \( \tilde{p}(\cdot) \) is a symmetric distribution satisfying \( \tilde{p}(\xi_t) = \tilde{p}(-\xi_t), \ \forall \xi_t \in \mathbb{R}^d \). Then gradient clipping with threshold \( c \) has the following properties:

1. If \( \|\nabla f(x_t)\| \leq \frac{3}{4} c \), then \( \mathbb{E}_{\xi_t \sim \tilde{p}}[\langle \nabla f(x_t), g_t \rangle] \geq \frac{3}{4} \|\nabla f(x_t)\| \mathbb{P}_{\xi_t \sim \tilde{p}}(\|\xi_t\| < \frac{c}{4}) \);
2. If \( \|\nabla f(x_t)\| > \frac{3}{4} c \), then \( \mathbb{E}_{\xi_t \sim \tilde{p}}[\langle \nabla f(x_t), g_t \rangle] \geq \frac{3}{4} \|\nabla f(x_t)\| \mathbb{P}_{\xi_t \sim \tilde{p}}(\|\xi_t\| < \frac{c}{4}) \).

Theorem 2 states that when the noise distribution is symmetric, gradient clipping will keep the expected clipped gradients positively aligned with the true gradient. This is the desired property that can guarantee convergence. Combining Theorem 2 with Theorem 1, we have Corollary 1 to fully characterize the convergence behavior of SGD with gradient clipping.

**Corollary 1.** Consider the SGD algorithm with gradient clipping given in (3). Set \( \alpha = \frac{1}{\sqrt{T}} \), and choose \( \tilde{p}(\cdot) \) as a symmetric distribution satisfying \( \tilde{p}_t(\xi_t) = \tilde{p}_t(-\xi_t), \ \forall \xi_t \in \mathbb{R}^d \). Then the following holds:

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_{\xi_t \sim \tilde{p}_t}(\|\xi_t\| < \frac{c}{4}) \min\left\{ \|\nabla f(x_t)\|, \frac{3}{4} c \right\} \|\nabla f(x_t)\| \leq \frac{D_f}{\sqrt{T}} + \frac{G}{2 \sqrt{T}} c^2 - \frac{1}{T} \sum_{t=1}^{T} b_t,
\]

where we have defined \( b_t := \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle (p_t(\xi_t) - \tilde{p}_t(\xi_t)) \rangle d\xi_t \).

The above result suggests that, as long as the probabilities \( \mathbb{P}_{\xi_t \sim \tilde{p}_t}(\|\xi_t\| < \frac{c}{4}) \) are bounded away from 0 and the symmetric distributions \( \tilde{p}_t \) are close approximations to \( p_t \) (small bias \( b_t \) then gradient norm goes to 0. Moreover, when \( \|\xi_t\| \) is drawn from a sub-gaussian distribution with constant variance, the probability does not diminish with the dimension. This is consistent with the observations in recent work of [Li et al. 2020], [Gur-Ari et al. 2018] on deep learning training, and we also provide our own empirical evaluation on the probability term in the Appendix. Note if the bias is negative and very large, the bound on the rhs will not be meaningful. Therefore, it is

\[1\] Both Theorem 2 and Corollary 1 hold under a more relaxed condition of \( \tilde{p}(\xi) = \tilde{p}(-\xi) \) for \( \xi \) with \( \ell_2 \) norm exceeding \( c/4 \).
useful to further study properties of such bias term. In the next section, we will discuss how large the bias term can be for a few choices of \( p \) and \( \bar{p} \). It turns out that the accumulation of \( b_t \) can help in some cases. In addition, one can extend the convergence results to some special non-symmetric distributions.

### 2.2 Beyond symmetric distributions

Theorem 2 and Corollary 1 suggest that as long as the distribution \( p \) is sufficiently close to a symmetric distribution \( \tilde{p} \), the convergence bias expressed as \( \sum_{t=1}^{T} b_t \) will be small. We now show that our bias decomposition result enables us to analyze the effect of the bias even for some highly asymmetric distributions. Note that when \( b_t \geq 0 \), the bias in fact helps convergence according Corollary 1.

We now provide three examples where \( b_t \) can be non-negative. Therefore, near-symmetricity is not a necessary condition for convergence, and our symmetricity-based analysis remains an effective tool to establish convergence for a broad class of distributions.

**Positively skewed.** Suppose \( p \) is positively skewed, that is, \( p(\xi) \geq p(-\xi) \), for all \( \xi \) with \( \langle \xi, \nabla f(x_t) \rangle > 0 \). With such distributions, the stochastic gradients tend to be positively aligned with the true gradient. If one chooses \( \tilde{p}(\xi_t) = 1/2 (p(\xi_t) + p(-\xi_t)) \), the bias \( b_t \) can be written as

\[
\int_{\xi_t \in \{\xi: \langle \xi, \nabla f(x_t) \rangle > 0\}} \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) - \text{clip}(\nabla f(x_t) - \xi_t, c) \rangle \left(\frac{1}{2} (p(\xi_t) - p(-\xi_t))\right) d\xi_t,
\]

which is always positive since \( \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) - \text{clip}(\nabla f(x_t) - \xi_t, c) \rangle \geq 0 \). Substituting into Theorem 2, we have \( \mathbb{E}_{\xi_t \sim \tilde{p}}[\langle \nabla f(x_t), g_t \rangle] \) strictly larger than \( \mathbb{E}_{\xi_t \sim p}[\langle \nabla f(x_t), g_t \rangle] \), which means the positive skewness helps the convergence (we want \( \mathbb{E}_{\xi_t \sim \tilde{p}}[\langle \nabla f(x_t), g_t \rangle] \) as large as possible).

**Mixture of symmetric.** The distribution of stochastic gradient \( \nabla f(x_t) + \xi_t \) is a mixture of two symmetric distributions \( p_0 \) and \( p_1 \) with mean 0 and \( v \) respectively. Such a distribution might be possible when most of samples are well classified. In this case, even though the distribution of \( \xi_t \) is not symmetric, one can apply similar argument of Theorem 2 to the component with mean \( v \), and the zero mean component yield a bias 0. In particular, let \( w_0 \) be the probability that \( \nabla f(x_t) + \xi_t \) is drawn from \( p_0 \). One can choose \( \bar{p} = p - w_0 p_0 \) which is the component symmetric over \( v \). The bias become

\[
\int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle w_0 p_0(\xi_t) d\xi_t = 0.
\]

This is because \( p_0(\xi_t) \) corresponds to a zero mean symmetric distribution of \( \nabla f(x_t) + \xi_t \), and the fact that if \( \nabla f(x_t) + \xi_t \) follows a symmetric distribution centered at 0, so does \( \text{clip}(\nabla f(x_t) + \xi_t, c) \) for any \( c > 0 \). Note that despite \( \bar{p} = p - w_0 p_0 \) is not a distribution since \( \int \bar{p}(\xi_t) = 1 - w_0 \), Theorem 2 can still be applied with everything on r.h.s. of inequalities multiplied by \( 1 - w_0 \) because one can apply Theorem 2 to distribution \( \bar{p}(\xi_t)/(1 - w_0) \) and then scale everything down.

**Mixture of symmetric or positively skewed.** If \( p \) is a mixture of multiple symmetric or positively skewed distributions, one can split the distributions into multiple ones and use their individual properties. That is, one can easily establish convergence guarantee for \( p \) being a mixture of \( m \) spherical distributions with mean \( u_1, ..., u_m \) and \( \langle f(x_t), u_i \rangle \geq 0, \forall i \in [m] \) as in the following theorem.
Theorem 5. Given $m$ distributions with the pdf of the $i$th distribution being $p_i(\xi) = \phi_i(\|\xi - u_i\|)$ for some function $\phi_i$. If $\nabla f(x_t) = \sum_{i=1}^{m} w_i u_i$ for some $w_i \geq 0, \sum_{i=1}^{m} w_i = 1$. Define a mixture of these distributions with zero mean as below:

$$p'(\xi) = \sum_{i=1}^{m} w_i p_i(\xi - \nabla f(x_t)).$$

If $\langle u_i, \nabla f(x_t) \rangle \geq 0, \forall i \in [m]$, we have

$$\mathbb{E}_{\xi_t \sim p'}[\langle \nabla f(x_t), g_t \rangle] \geq \|\nabla f(x_t)\| \sum_{i=1}^{m} w_i \min\left(\|u_i\|, \frac{3}{4} c\right) \cos(\nabla f(x_t), u_i) \mathbb{P}_{\xi_t \sim p_i} \left(\|\xi_t\| < \frac{c}{4}\right) \geq 0.$$

Besides these examples of favorable biases above, there are also many cases where $b_t$ can be negative and lead to a convergence gap, such as negatively skewed distributions or multimodal distributions with highly imbalanced modes. We have illustrated possible distributions in our divergence examples (Examples 1 and 2). In such cases, one should expect that clipping has an adversarial impact on the convergence guarantee. However, as we also show in Section 4, the gradient distributions on real datasets tend to be symmetric, so their clipping biases are small.

3 DP-SGD with Gradient Clipping

We now extend the results above to analyze the overall convergence DP-SGD with gradient clipping. To match up with the setting in Section 2, we consider the distribution $D$ to be the empirical distribution over a private dataset $S$ of $n$ examples $\{s_1, \ldots, s_n\}$, and so $f(x) = \frac{1}{n} \sum_{i=1}^{n} f(x, s_i)$. For any iterate $x_t \in \mathbb{R}^d$ and example $s_i$, let $\xi_{t,i} = \nabla f(x_t, s_i) - \nabla f(x_t)$ denote the gradient noise on the example, and $p_t$ denote the distribution over $\xi_{t,i}$. At each iteration $t$, DP-SGD performs:

$$x_{t+1} = x_t - \alpha \left(\frac{1}{|S_t|} \sum_{i \in S_t} \text{clip}(\nabla f(x_t) + \xi_{t,i}, c) + Z_t\right),$$

where $S_t$ is a random subsample of $S$ (with replacement) and $Z_t \sim \mathcal{N}(0, \sigma^2 I)$ is the noise added for privacy. We first recall the privacy guarantee of the algorithm below:

Theorem 4 (Privacy (Theorem 1 in [Abadi et al., 2016b])). There exist constants $u$ and $v$ so that given the number of iterations $T$, for any $\epsilon \leq u q^2 T$, where $q = \frac{|S|}{n}$, DP-SGD with gradient clipping of threshold $c$ is $(\epsilon, \delta)$-differentially private for any $\delta > 0$, if $\sigma^2 \geq \frac{c^2 T \ln(\frac{1}{\delta})}{n^2 e^2}.$

By accounting for the sub-sampling noise and Gaussian perturbation in DP-SGD, we obtain the following convergence guarantee, where we further bound the clipping bias term $b_t$ with the Wasserstein distance between the gradient distribution and a coupling symmetric distribution.

Theorem 5 (Convergence). Suppose $x \in \mathbb{R}^d$, let $m = |S|$, and let $\tilde{p}_t$ be a symmetric distribution with $\tilde{p}_t(\xi_t) = \tilde{p}_t(-\xi_t), \forall \xi_t \in \mathbb{R}^d$. For DP-SGD with gradient clipping, set

$$\alpha = \frac{\sqrt{D f \ln(\frac{1}{\delta})}}{n c \epsilon L}.$$
Then there exist \( u \) and \( v \) such that for any \( \varepsilon \leq u \frac{\ln^2 T}{n^2}, \sigma^2 = v \frac{c^2 T \ln(\frac{1}{\varepsilon})}{n^2 \varepsilon^2} \), we have

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_{\xi_t \sim \tilde{P}} \left( \| \xi_t \| < \frac{c}{4} \right) h_c(\| \nabla f(x_t) \|) \leq \frac{1}{2} v + \frac{3}{2} c \frac{\sqrt{D_f G_d \ln(\frac{1}{\varepsilon})}}{n \varepsilon} + \frac{1}{T} \sum_{t=1}^{T} W_{\tilde{f}(x_t), \tilde{P}}(\tilde{p}_t, p_t),
\]

where \( h_c(y) := \min(y^2, \frac{3}{4} cy) \); \( D_f := f(x_1) - \min_x f(x) \); \( W_{v,c}(p, p') \) is the Wasserstein distance between \( p \) and \( p' \) with metric function

\[
d_{v,c}(a, b) := |\langle v, \text{clip}(v + a, c) \rangle - \langle v, \text{clip}(v + b, c) \rangle|.
\]

**Remark on the Wasserstein distance.** In (6), it is clear that the convergence bias \( b_t \) can be bounded by the total variation distance between \( p_t \) and \( \tilde{p}_t \) or some similar distance between distributions such as the one below

\[
b_t = \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle (\tilde{p}_t(\xi_t) - p_t(\xi_t)) d\xi_t
\]

\[
\leq c \cdot \| \nabla f(x_t) \| \int |p_t(\xi_t) - \tilde{p}_t(\xi_t)| d\xi_t.
\]

(9)

However, the above bound (or the total variation distance) becomes trivial when \( p_t \) is the empirical distribution over a finite sample, because it is always 2 (always 1 for the total variation distance) when \( \tilde{p} \) is continuous. In addition, the bias is hard to interpret without further transformation. This is why we bound \( b_t \) by the Wasserstein distance as follows:

\[
b_t = \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle (\tilde{p}_t(\xi_t) - p_t(\xi_t)) d\xi_t
\]

\[
= \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle \tilde{p}_t(\xi_t) d\xi_t - \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t', c) \rangle p_t(\xi_t') d\xi_t'
\]

\[
= \int \int ((\langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle - \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t', c) \rangle) \gamma(\xi_t, \xi_t') d\xi_t d\xi_t'
\]

\[
\leq \int \int |\langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle - \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t', c) \rangle| \gamma(\xi_t, \xi_t') d\xi_t d\xi_t',
\]

(10)

where \( \gamma(\cdot, \cdot) \) is any joint distribution with marginal \( \tilde{p}(\cdot) \) and \( p(\cdot) \). Thus, we have

\[
-b_t \leq \inf_{y \in \Gamma(\tilde{p}, p)} \int |\langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle - \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t', c) \rangle| \gamma(\xi_t, \xi_t') d\xi_t d\xi_t'.
\]

where \( \Gamma(\tilde{p}, p) \) is the set of all couplings with marginals \( \tilde{p} \) and \( p \) on the two factors, respectively. If we define the distance function

\[
d_{\gamma,c}(a, b) := |\langle y, \text{clip}(y + a, c) \rangle - \langle y, \text{clip}(y + b, c) \rangle|.
\]

Then we have

\[
b_t \leq \inf_{y \in \Gamma(\tilde{p}, p)} \int d_{\gamma,f(x_t),c}(\xi_t, \xi_t') \gamma(\xi_t, \xi_t') d\xi_t d\xi_t'.
\]

(11)

The right-hand side (r.h.s.) of the above inequality is the Wasserstein distance defined on the distance function \( d_{\gamma,f(x_t),c} \). It converges to the distance between the population distribution of
gradient and \( \tilde{p} \) with \( n \) being large since the empirical distribution will be similar to the population distribution.

Thus, if the population distribution of gradient is approximate symmetric, the bias term tends to be small. In addition, the distance function is uniformly bounded by \( \| \nabla f(x) \| c \) which makes it is more favorable than \( \ell_2 \) distance. Compared with the expression of \( b_t \) in Corollary \( \text{[1]} \) the Wasserstein distance is easier to interpret when \( \tilde{p} \) is discrete.

### 4 Experiments

In this section, we investigate whether the gradient distributions of DP-SGD are approximate symmetric in practice. However, since the gradient distributions are high-dimensional, certifying symmetricity is in general intractable. We instead consider two simple proxy measures and visualizations.

**Setup.** We run DP-SGD implemented in Tensorflow\(^3\) on two popular datasets MNIST [LeCun et al., 2010] and CIFAR-10 [Krizhevsky et al., 2009]. For MNIST, we train a CNN with two convolution layers with 16 4\times 4 kernels followed by a fully connected layer with 32 nodes. We use DP-SGD to train the model with \( \alpha = 0.15 \), and a batchsize of 128. For CIFAR-10, we train a CNN with two convolutional layers with 2\times 2 max pooling of stride 2 followed by a fully connected layer, all using ReLU activation, each layer uses a dropout rate of 0.5. The two convolution layers respectively has 32 and 64 kernels, each of size 3\times 3; further the fully connected layer has 1,500 nodes. We use \( \alpha = 0.001 \) and decrease it by 10 times every 20 epochs. The clip norm of both experiments is set to be \( c = 1 \) and the noise multiplier is 1.1.

**Visualization with random projections.** We visualize the gradient distribution by projecting the gradient to a two-dimensional space using random Gaussian matrices. Note that given any symmetric distribution, its two-dimensional projection remains symmetric for any projection matrix. On the contrary, if for all projection matrix, the projected gradient distribution is symmetric, the original gradient distribution should also be symmetric. We repeat the projection using different randomly generated matrices and visualize the induced distributions.

We can see that on both datasets, the gradient distribution is non-symmetric before training (Epoch 0), but over the epochs, the gradient distributions become increasingly symmetric. The distribution of gradients on MNIST at the end of epoch 9 projected to a random two-dimensional space using different random matrices is shown in Figure \( \text{[2]} \). It can be seen that the approximate symmetric property holds for all 8 realizations. We provide many more visualizations from different realized random projections across different epochs in the Appendix.

**Symmetry of angles.** We also measure the cosine similarities between per-sample stochastic gradients and the true gradient. We observe that the cosine similarities between per-sample stochastic gradients and the true gradient, defined as \( \cos(\nabla f(x_t) + \xi_{t,i}, \nabla f(x_t)) \), is approximate symmetric around 0 as shown in the histograms in Figure \( \text{[3]} \).

\(^3\)https://github.com/tensorflow/privacy/tree/master/tutorials
Figure 1: Gradient distributions on MNIST (top row) and CIFAR10 (bottom row) at the end of different epochs (indexed by columns). The gradients for epoch 0 are computed at initialization (before training).

Figure 2: Gradient distributions on MNIST at the end of epoch 9 projected using different random matrices.

Figure 3: Histogram of cosine between stochastic gradients and the true gradient at the end of different epochs.
5 Mitigating Clipping Bias with Perturbation

From previous analyses, SGD with gradient clipping and DP-SGD have good convergence performance when the gradient noise distribution is approximately symmetric or when the gradient bias favors convergence (e.g., mixture of symmetric distributions with aligned mean). Although in practice, gradient distributions do exhibit (approximate) symmetry (see Sec. 4), it would be desirable to have tools to handle situations where the clipping bias does not favor convergence. Now we provide an approach to decrease the bias. If one adds some Gaussian noise before clipping, i.e.

\[ g_t = \text{clip}(\nabla f(x_t) + \xi_t + k\zeta_t, c), \zeta_t \sim \mathcal{N}(0, I) \]

we can prove \(|b_t| = O\left(\frac{\sigma^2}{k^2}\right)\) as in Theorem 6.

**Theorem 6.** Let \( g_t = \text{clip}(\nabla f(x_t) + \xi_t + k\zeta_t, c) \) and \( \zeta_t \sim \mathcal{N}(0, I) \). Then gradient clipping algorithm has following properties:

\[
\mathbb{E}_{\xi_t \sim p, \zeta_t} [\langle \nabla f(x_t), g_t \rangle] \geq \|\nabla f(x_t)\| \min \left\{ \|\nabla f(x_t)\|, \frac{3}{4}c \right\} P(\|k\zeta_t\| < c) - \|\nabla f(x_t)\| O\left(\frac{\sigma^2_{\xi_t}}{k^2}\right),
\]

where \( \sigma^2_{\xi_t} \) is the variance of the gradient noise \( \xi_t \).

More discussion can be found in the Appendix. By adding the noise, one trades off bias with variance. Larger noise makes the algorithm converges possibly slower but better. This trick can be helpful when the gradient distribution is not favorable. To verify its effect in practice, we run SGD with gradient clipping on a few unfavorable problems including examples in Section 1 and a new high dimensional example. For the new example, we minimize the function \( f(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \|x - z_i\|^2 \) with \( n = 10000 \). Each \( z_i \) is drawn from a mixture of isotropic Gaussian with 3 components of dimension 10. The covariance matrix of all components is \( I \) and the means of the 3 components are drawn from \( \mathcal{N}(0, 36I), \mathcal{N}(0, 4I), \mathcal{N}(0, I) \), respectively. We set \( \alpha = 0.1 \) for the new examples and \( \alpha = 0.001 \) for the examples in Section 1. Figure 5 shows \( \|x_t - \arg\min_x f(x)\| \) versus \( t \). We can see that SGD with gradient clipping converges to non-optimal points as predicted by theory. In contrast, pre-clipping perturbation ensures convergence.

6 Conclusion

In this paper, we provide a theoretical analysis on the effect of gradient clipping in SGD and private SGD. We provide a new way to quantify the clipping bias by coupling the gradient distribution with a geometrically symmetric distribution. Combined with our empirical evaluation
showing that gradient distribution of private SGD follows some symmetric structure along the trajectory, these results provide an explanation why gradient clipping works in practice. We also provide a perturbation-based technique to reduce the clipping bias even for adversarial instances.
References

Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derekordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. CoRR, abs/1603.04467, 2016a. URL http://arxiv.org/abs/1603.04467.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pages 308–318, 2016b.

Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. pages 464–473, 2014. doi: 10.1109/FOCS.2014.56. URL https://doi.org/10.1109/FOCS.2014.56.

Brett K. Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P. Bhavnani, James Brian Byrd, and Casey S. Greene. Privacy-preserving generative deep neural networks support clinical data sharing. Circulation: Cardiovascular Quality and Outcomes, 12(7):e005122, 2019. doi: 10.1161/CIRCOUTCOMES.118.005122. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCOUTCOMES.118.005122.

Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. Deep learning with gaussian differential privacy. arXiv preprint arXiv:1911.11607, 2019.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography Conference, pages 265–284. Springer, 2006.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. CoRR, abs/1812.04754, 2018. URL http://arxiv.org/abs/1812.04754.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2, 2010.

Xinyan Li, Qilong Gu, Yingxue Zhou, Tiancong Chen, and Arindam Banerjee. Hessian based analysis of SGD for deep nets: Dynamics and generalization. In Carlotta Damiani and Nitesh V. Chawla, editors, Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020, Cincinnati, Ohio, USA, May 7-9, 2020 [the conference was canceled because of the coronavirus pandemic, the reviewed papers are published in this volume], pages 190–198. SIAM, 2020. doi: 10.1137/1.9781611976236.22. URL https://doi.org/10.1137/1.9781611976236.22.

Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. arXiv preprint arXiv:1908.07643, 2019.
Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In 2013 IEEE Global Conference on Signal and Information Processing, pages 245–248. IEEE, 2013.

Om Thakkar, Galen Andrew, and H. Brendan McMahan. Differentially private learning with adaptive clipping. CoRR, abs/1905.03871, 2019. URL http://arxiv.org/abs/1905.03871.

Di Wang and Jinhui Xu. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 1182–1189. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33011182. URL https://doi.org/10.1609/aaai.v33i01.33011182.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 1226–1235, 2019.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In International Conference on Learning Representations, 2019.

Yuqing Zhu and Yu-Xiang Wang. Poisson subsampled rényi differential privacy. In International Conference on Machine Learning, pages 7634–7642, 2019.
A Proof of Theorem 1

By smoothness assumption, we have

\[ f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{G}{2} \|x_{t+1} - x_t\|^2. \] (14)

Then, by update rule and the fact that \( \|g_t\| \leq c \), we have

\[ f(x_{t+1}) \leq f(x_t) - \alpha \langle \nabla f(x_t), g_t \rangle + \frac{G\alpha^2 c^2}{2}. \] (15)

Take expectation, sum over \( t \) from 1 to \( T \), divide both sides by \( T\alpha \), rearranging and substituting into \( \alpha = \frac{1}{\sqrt{T}} \), we get

\[ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\langle \nabla f(x_t), g_t \rangle] \leq \frac{1}{T\alpha} (f(x_1) - f(x_{T+1})) + \frac{G\alpha c^2}{2} \]

\[ \leq \frac{1}{\sqrt{T}} \mathbb{E}[f(x_1) - f(x_{T+1})] + \frac{Gc^2}{2\sqrt{T}} \]

\[ \leq \frac{1}{\sqrt{T}} D_f + \frac{Gc^2}{2\sqrt{T}}, \] (16)

where \( D_f = f(x_1) - \min_x f(x) \).

\[ \square \]

B Proof of Theorem 2

In the proof, we assume \( \xi_t \sim \tilde{p}_t \), and we omit subscript of \( P \) and \( \mathbb{E} \) to simplify notations. Further, we will denote \( \bar{g}_t \triangleq \nabla f(x_t) \).

B.1 When gradient is small

Let us first consider the case with \( \|\bar{g}_t\| \leq \frac{3}{4} c \).

Denote \( B \) to be the event that \( \|\bar{g}_t + \xi_t\| \leq c \) and \( \|\bar{g}_t - \xi_t\| \leq c \), we have \( \mathbb{P}(B) \geq \mathbb{P}(\|\xi_t\| \leq \frac{c}{4}) \). Define \( D = \{\xi : \|\bar{g}_t + \xi_t\| > c \text{ or } \|\bar{g}_t - \xi_t\| > c\} \). Taking an expectation conditioning on \( x_t \), we have

\[ \mathbb{E}[\langle \bar{g}_t, g_t \rangle] = \langle \bar{g}_t, \mathbb{E}[\text{clip}(\bar{g}_t + \xi_t, c)] \rangle \]

\[ = \langle \bar{g}_t, \mathbb{E}\left[ \text{clip}(\bar{g}_t + \xi_t, c) \mathbb{I}(B) + \text{clip}(\bar{g}_t + \xi_t, c) \mathbb{I}(D) \right] \rangle \]

\[ \geq \|\bar{g}_t\|^2 \mathbb{P}(\|\xi_t\| \leq \frac{c}{4}) + \langle \bar{g}_t, \int_D \text{clip}(\bar{g}_t + \xi_t, c) \hat{p}(\xi_t) d\xi_t \rangle, \]

where the last inequality is due to \( \text{clip}(\bar{g}_t + \xi_t, c) = \bar{g}_t + \xi_t \) when \( B \) happens and \( \mathbb{P}(B) \geq \mathbb{P}(\|\xi_t\| \leq \frac{c}{4}) \) and the assumption that \( \hat{p}(\xi) = \hat{p}(-\xi) \).
Now we need to analyze $T_1$. We have

$$T_1 = \frac{1}{2} \left( \int_D \text{clip}(\mathbf{g}_t + \mathbf{c}, c) \rho(\mathbf{c}_t) d\mathbf{c}_t + \int_D \text{clip}(\mathbf{g}_t - \mathbf{c}, c) \rho(\mathbf{c}_t) d\mathbf{c}_t \right)$$

$$= \frac{1}{2} \left( \int_D \left( \text{clip}(\mathbf{g}_t + \mathbf{c}, c) + \text{clip}(\mathbf{g}_t - \mathbf{c}, c) \right) \rho(\mathbf{c}_t) d\mathbf{c}_t \right)$$

$$= \frac{1}{2} \|\tilde{g}_t\| \times \int_D \left( \|\text{clip}(\mathbf{g}_t + \mathbf{c}, c)\| \cos(\mathbf{g}_t, \mathbf{g}_t + \mathbf{c}) + \|\text{clip}(\mathbf{g}_t - \mathbf{c}, c)\| \cos(\mathbf{g}_t, \mathbf{g}_t - \mathbf{c}) \right) \rho(\mathbf{c}_t) d\mathbf{c}_t, \tag{17}$$

where the last equality is because $\langle a, b \rangle = \|a\| \|b\| \cos(a, b)$ for any vector $a, b$, and that the clipping operation keeps directions.

Now we will show that $T_2(\mathbf{c}_t) \geq 0$. Towards this end, let us consider three cases.

**Case I.** Suppose $\|\mathbf{g}_t + \mathbf{c}_t\| \geq c$ and $\|\mathbf{g}_t - \mathbf{c}_t\| \geq c$. In this case, we have

$$T_2(\mathbf{c}_t) = c \cdot (\cos(\mathbf{g}_t, \mathbf{g}_t + \mathbf{c}) + \cos(\mathbf{g}_t, \mathbf{g}_t - \mathbf{c})) \geq 0, \tag{18}$$

where the inequality is due to Lemma 1.

**Case II.** One of $\|\mathbf{g}_t + \mathbf{c}_t\|$ and $\|\mathbf{g}_t - \mathbf{c}_t\|$ is less than $c$.

**Case II (a).** First we assume $\cos(\mathbf{g}_t, \mathbf{g}_t + \mathbf{c}) < 0$. Then we must have $\cos(\mathbf{g}_t, \mathbf{g}_t - \mathbf{c}) < 0$ so that $\langle \mathbf{g}_t, -\mathbf{c} \rangle < 0$ and $\cos(\mathbf{g}_t, \mathbf{g}_t + \mathbf{c}) = \frac{\langle \mathbf{g}_t, \mathbf{g}_t + \mathbf{c} \rangle}{\|\mathbf{g}_t\| \|\mathbf{g}_t + \mathbf{c}\|} > 0$. Then, from Lemma 1, we have

$$\|\mathbf{g}_t + \mathbf{c}\| \geq \|\mathbf{g}_t - \mathbf{c}\|, \tag{19}$$

and it follows that $\|\mathbf{g}_t - \mathbf{c}\| \leq c$. So in this case, we have

$$T_2(\mathbf{c}_t) = \|\text{clip}(\mathbf{g}_t + \mathbf{c}, c)\| \cos(\mathbf{g}_t, \mathbf{g}_t + \mathbf{c}) + \|\text{clip}(\mathbf{g}_t - \mathbf{c}, c)\| \cos(\mathbf{g}_t, \mathbf{g}_t - \mathbf{c})$$

$$= c \cdot \cos(\mathbf{g}_t, \mathbf{g}_t + \mathbf{c}) + \|\text{clip}(\mathbf{g}_t - \mathbf{c}, c)\| \cos(\mathbf{g}_t, \mathbf{g}_t - \mathbf{c})$$

$$\geq c \cdot \cos(\mathbf{g}_t, \mathbf{g}_t + \mathbf{c}) + c \cdot \cos(\mathbf{g}_t, \mathbf{g}_t - \mathbf{c}) \geq 0, \tag{20}$$

where the last inequality is due to Lemma 1.

**Case II (b).** Similar argument applies to the case with $\cos(\mathbf{g}_t, \mathbf{g}_t + \mathbf{c}) < 0$.

**Case II (c).** Suppose $\cos(\mathbf{g}_t, \mathbf{g}_t + \mathbf{c}) \geq 0, \cos(\mathbf{g}_t, \mathbf{g}_t - \mathbf{c}) \geq 0$. Then $T_2(\mathbf{c}_t) \geq 0$ holds trivially.

In summary, we have shown that the following holds:

$$\mathbb{E}[\langle \mathbf{g}_t, \mathbf{g}_t \rangle] \geq \|\mathbf{g}_t\|^2 \mathbb{P}(\|\mathbf{c}_t\| \leq \frac{c}{4}). \tag{21}$$

This completes the proof.

**B.2 When gradient is large**

Now let us look at the case where gradient is large, i.e. $\|\mathbf{g}_t\| \geq \frac{3}{4} c$. 

By definition, we have

\[
\mathbb{E}[\langle \vec{g}_t, \vec{g}_t \rangle] = \int_{\xi_t} \langle \vec{g}_t, \text{clip}(\vec{g}_t + \xi, c) \cdot p(\xi_t) \rangle d\xi_t
\]

\[
= \int_{\xi_t} \langle \vec{g}_t, \text{clip}(\vec{g}_t + \xi, c) \rangle \cdot p(\xi_t) d\xi_t
\]

\[
= \|\vec{g}_t\| \int_{\xi_t} \|\text{clip}(\vec{g}_t + \xi, c)\| \cos(\langle \vec{g}_t, \text{clip}(\vec{g}_t + \xi, c) \rangle) \cdot p(\xi_t) d\xi_t
\]

\[
(i) = \|\vec{g}_t\| \int_{\xi_t} \|\text{clip}(\vec{g}_t + \xi, c)\| \cos(\langle \vec{g}_t, \text{clip}(\vec{g}_t + \xi, c) \rangle) \cdot p(\xi_t) d\xi_t
\]

\[
(ii) = \|\vec{g}_t\| \int_{\xi_t} \|\text{clip}(\vec{g}_t + \xi, c)\| \cos(\langle \vec{g}_t, \text{clip}(\vec{g}_t + \xi, c) \rangle) \cdot p(\xi_t) d\xi_t,
\]

where in (i) we used the fact that clipping operation keeps the direction, that is:

\[
\frac{\text{clip}(\vec{g}_t + \xi, c)}{\|\text{clip}(\vec{g}_t + \xi, c)\|} = \frac{\vec{g}_t + \xi}{\|\vec{g}_t + \xi\|}, \quad \forall \xi, \forall c > 0.
\]

In (ii) we have introduced an arbitrary rotation matrix \(G\), and we have used the fact that the angle between two vectors remains the same after the same rotation, and that the norm of the clip operation is rotation invariant:

\[
\|\text{clip}(Gz, c)\| = \|\text{clip}(z, c)\|, \quad \forall z \in \mathbb{R}^d, \forall c > 0.
\]

In the following, we will show \(T_3\) is a non-decreasing function of \(\vec{g}_t\).

Since the rotation matrix \(G\) in \(T_3\) is arbitrary, without loss of generality (wlog), we can assume that the first element of \(\vec{g}_t\) equals \(\|\vec{g}_t\|\) and the rest are all zeros, that is:

\[
\vec{g}_t[1] = \|\vec{g}_t\| > 0, \quad \vec{g}_t[i] = 0, \quad 2 \leq i \leq d.
\]

For notation simplicity, let us define \(\vec{y} := \|\vec{g}_t\|\). Then to show \(T_3\) is a non-decreasing function of \(\|\vec{g}_t\|\), it is sufficient to show that each term in the integration is a non-decreasing function of \(\vec{y}\). That is, for all \(\xi_t\), the following quantity is a non-decreasing function of \(\vec{y}\) for \(\vec{y} > 0\) when \(\vec{g}_t = [\vec{y}, 0, ..., 0]\):

\[
\|\text{clip}(\vec{g}_t + \xi, c)\| \cos(\langle \vec{g}_t, \text{clip}(\vec{g}_t + \xi, c) \rangle).
\]

To this end, we divide our analysis into two cases.

**Case I:** Suppose \(\|\vec{g}_t + \xi_t\| \leq c\). In this case, (25) reduces to

\[
\|\vec{g}_t + \xi_t\| \cos(\langle \vec{g}_t, \vec{g}_t + \xi_t \rangle) = \|\vec{g}_t + \xi_t\| \frac{\langle \vec{g}_t, \vec{g}_t + \xi_t \rangle}{\|\vec{g}_t + \xi_t\|}
\]

\[
= \frac{\langle \vec{g}_t, \vec{g}_t + \xi_t \rangle}{\|\vec{g}_t\|} \frac{\vec{y}(\vec{y} + \xi_{t,1})}{\vec{y}} = \vec{y} + \xi_{t,1}.
\]

Clearly, the above quantity is a monotonically increasing function of \(\vec{y}\).
Case II: Suppose $\|\tilde{g}_t + \xi_t\| \geq c$. Then we have

$$
\|\text{clip}(\tilde{g}_t + \xi_t, c)\| \cos(\tilde{g}_t, \tilde{g}_t + \xi_t) = c \cdot \cos(\tilde{g}_t, \tilde{g}_t + \xi_t) = c \frac{\langle \tilde{g}_t, \tilde{g}_t + \xi_t \rangle}{\|\tilde{g}_t\|\|\tilde{g}_t + \xi_t\|} = c \frac{y(y + \xi_{t,1})}{y \sqrt{(y + \xi_{t,1})^2 + \sum_{i=2}^d \xi_{t,i}^2}} = c \frac{(y + \xi_{t,1})}{\sqrt{(y + \xi_{t,1})^2 + \sum_{i=2}^d \xi_{t,i}^2}}.
$$

(27)

It is also easy to verify that the above function is a non-decreasing function of $y$.

To see it is non-decreasing, define

$$r(z) = c \frac{z}{\sqrt{z^2 + q^2}}, \text{ with } c > 0.
$$

(28)

Then it is easy to check that $r'(z) = c(1 - \frac{z^2}{z^2 + q^2}) \geq 0$. By letting $z = y + \xi_{t,1}$ and $q^2 = \sum_{i=2}^d \xi_{t,i}^2$, we conclude that the r.h.s. of (27) is non-decreasing.

Since the clipping function is continuous, combined with the above results, we know (25) is a non-decreasing function of $\|\tilde{g}_t\|$. Then by utilizing the above non-decreasing property, and the assumption that $\|\tilde{g}_t\| \geq 3c/4$, we have the following

$$
E[\langle \tilde{g}_t, \xi_t \rangle] = E[\langle \tilde{g}_t, \tilde{g}_t + \xi_t \rangle] \\
= \|\tilde{g}_t\| T_3 \geq \|\tilde{g}_t\| T_3 \\
\geq \|\tilde{g}_t\| \int_{\xi_t} \left\|\text{clip}\left(3 \cdot \frac{c}{4} \frac{\tilde{g}_t}{\|\tilde{g}_t\|}, + \xi_t, c\right)\right\| \cos\left(3 \cdot \frac{c}{4} \frac{\tilde{g}_t}{\|\tilde{g}_t\|}, 3 \cdot \frac{c}{4} \frac{\tilde{g}_t}{\|\tilde{g}_t\|} + \xi_t\right) \cdot p(\xi_t) d\xi_t \\
= E\left[\left(\frac{3}{4} \cdot \frac{c}{\|\tilde{g}_t\|}, \frac{3}{4} \cdot \frac{\tilde{g}_t}{\|\tilde{g}_t\|} + \xi_t\right)\right] \\
= E[\langle \tilde{g}_t, \tilde{g}_t + \xi_t \rangle],
$$

(29)

where we have defined $\tilde{g}_t := \frac{3c}{4} \frac{\tilde{g}_t}{\|\tilde{g}_t\|}$, with $\|\tilde{g}_t\| = \frac{3}{4} c$.

From the first part of the theorem, we know for any vector $\|z\| = \frac{3}{4} c$, the following holds

$$
E[\langle z, \xi_t \rangle] \geq \|z\|^2 p\left(\|\xi_t\| < \frac{c}{4}\right) = \|z\|^{2} p\left(\|\xi_t\| < \frac{c}{4}\right).
$$

(30)

Combining the above result with (29), we obtain the following:

$$
\|\tilde{g}_t\| T_3 \geq E[\langle \tilde{g}_t, \tilde{g}_t \rangle] \geq \|\tilde{g}_t\| p\left(\|\xi_t\| < \frac{c}{4}\right).
$$

(31)

This implies $T_3 \geq \frac{3}{4} c \cdot p\left(\|\xi\| < \frac{c}{4}\right)$. So we obtain

$$
E[\langle \tilde{g}_t, \tilde{g}_t \rangle] = E[\langle \tilde{g}_t, \tilde{g}_t + \xi_t \rangle] = \|\tilde{g}_t\| T_3 \geq \|\tilde{g}_t\| \frac{3}{4} c \cdot p\left(\|\xi\| < \frac{c}{4}\right).
$$

(32)

The proof is completed.
B.3 Technical lemmas

Lemma 1. For any \( g \) and \( \xi \), we have

\[
\cos(g, g + \xi) + \cos(g, g - \xi) \geq 0.
\]

Proof: By definition of the cosine function, we have

\[
\cos(g, g + \xi) + \cos(g, g - \xi) = \frac{\langle g, g + \xi \rangle + \langle g, g - \xi \rangle}{\|g\|\|g + \xi\| + \|g\|\|g - \xi\|}
\]

where in the last equality we have defined \( e = \cos(g, \xi) \).

To prove the desired result, we only need the numerator of r.h.s. of (33) to be non-negative. Denote \( h(\xi) = \cos(g, g + \xi) + \cos(g, g - \xi) \), since \( h \) is rotation invariant, we can assume wlog that \( \xi_1 = \|\xi_1\| > 0 \) and \( \xi_{i, i} = 0 \) for \( 2 \leq i \leq d \). Also, because \( h(\xi) = h(-\xi) \), we can assume wlog that \( g_1 \geq 0 \).

Now suppose \( g_1 = a > 0 \), \( \sum_{i=2}^{d} g_i^2 = b^2 \). Denote the numerator of r.h.s. of (33) as \( T_4 \), it can be written as

\[
T_4 = \|g + \xi\|(\|g\| - \|\xi\|\|e\|) + \|g - \xi\|(\|g\| + \|\xi\|\|e\|)
\]

where we have defined

\[
e := \cos(g, \xi) = \frac{\langle g, \xi \rangle}{\|g\|\|\xi\|} = \frac{a}{\sqrt{a^2 + b^2}} \geq 0.
\]

Now let us analyze the sign of \( T_4 \). Recall that by assumption, \( \xi_1 > 0 \) and \( e \geq 0 \). Then we know \( T_6 \geq 0 \). We have \( T_4 \geq 0 \) trivially when \( T_5 \geq 0 \), i.e. when \( \xi_1 e \leq \sqrt{a^2 + b^2} \).

Now assume \( \xi_1 e > \sqrt{a^2 + b^2} \). Below we will show that \( T_6^2 \geq T_5^2 \), which implies that \( T_4 \geq 0 \). To this end, have we calculate the differences of \( T_6^2 \) and \( T_5^2 \) as:

\[
T_6^2 - T_5^2 = (b^2 + (a - \xi_1)^2)(\sqrt{a^2 + b^2} + \xi_1 e)^2 - (b^2 + (a + \xi_1)^2)(\sqrt{a^2 + b^2} - \xi_1 e)^2
\]

\[
= 4b^2 \xi_1 e \sqrt{a^2 + b^2} + 4\xi_1 e \sqrt{a^2 + b^2} (a^2 + \xi_1^2) - 4a \xi_1 (a^2 + b^2 + \xi_1^2 e^2).
\]
For $T_7$, we can further simplify it as

$$T_7 = 4\xi_1 e \sqrt{a^2 + b^2(a^2 + \xi_1^2)} - 4a\xi_1(a^2 + b^2 + \xi_1^2 e^2)$$

Combining all above, we have

$$T_7 \geq 0 \implies T_6^2 - T_5^2 \geq 0 \implies T_4 \geq 0 \implies \cos(g, g + \xi) + \cos(g, g - \xi) \geq 0.$$ 

This completes the proof.

**Lemma 2.** For any $g$ and $\xi$, we have:

$$\|g + \xi\| \geq \|g - \xi\|, \quad \text{if} \quad \cos(g, \xi) > 0,$$

$$\|g + \xi\| \leq \|g - \xi\|, \quad \text{if} \quad \cos(g, \xi) < 0.$$  \hspace{1cm} (35a) (35b)

**Proof:** Let us express $\xi$ using a coordinate system with one axis parallel to $g$. Define the basis of this coordinate system as $v_1, v_2, ... v_d$ with $v_1 = g/\|g\|$. Then we have $\xi = \sum_{i=1}^d b_i v_i$ and $\cos(g, \xi) > 0$ if and only if $b_1 > 0$. In addition, we have

$$\|g + \xi\| = \sqrt{\|g\|^2 + b_1^2 + 2\|g\|b_1 + \sum_{i=2}^d b_i^2},$$

and

$$\|g - \xi\| = \sqrt{\|g\|^2 - b_1^2 + 2\|g\|b_1 + \sum_{i=2}^d b_i^2}.$$ 

Then it is clear that $\|g + \xi\| \geq \|g - \xi\|$ when $b_1 > 0$. This completes the proof of of (35a). Similar arguments applies to the case with $\cos(g, \xi) < 0$.

---

**C Proof of Theorem 3**

**Theorem 3.** Given $m$ distributions with the pdf of the $i$th distribution being $p_i(\xi_i) = \phi_i(\|\xi_i - u_i\|)$ for some function $\phi_i$. If $\nabla f(x_i) = \sum_{i=1}^m w_i u_i$ for some $w_i \geq 0, \sum_{i=1}^m w_i = 1$. Let $p'(\xi_i) = \sum_{i=1}^m w_i p_i(\xi_i - \nabla f(x_i))$, be a mixture of these distributions with zero mean. If $(u_i, \nabla f(x_i)) \geq 0, \forall i \in [m]$, we have:

$$\mathbb{E}_{\xi_i \sim p'}[\langle \nabla f(x_i), g_i \rangle] \geq \|\nabla f(x_i)\| \sum_{i=1}^m w_i \min \left(\|u_i\|, \frac{3}{4}c\right) \cos(\nabla f(x_i), u_i) \mathbb{E}_{\xi_i \sim p'}[\|\xi_i\| < \frac{c}{4}] \geq 0.$$
**Proof:** First, we notice that Theorem 2 can be restated into a more general, which holds for any vector \( g \) instead of \( \nabla f(x_i) \), as follows.

**Theorem 2** (restated). Given a random variable \( \xi \sim \hat{\rho} \) with \( \hat{\rho}(\xi) \) being a symmetric distribution. Then for any vector \( g \in \mathbb{R}^d \), we have

1. If \( \|g\| \leq \frac{3}{4}c \), then  
   \[ \mathbb{E}[\langle g, \text{clip}(g + \xi, c) \rangle] \geq \|g\|^2 \mathbb{P}(\|\xi\| < \frac{c}{4}); \]  

2. If \( \|g\| > \frac{3}{4}c \), then  
   \[ \mathbb{E}[\langle g, \text{clip}(g + \xi, c) \rangle] \geq \frac{3}{4}c \|g\| \mathbb{P}(\|\xi\| < \frac{c}{4}). \]  

In addition, if \( \xi \sim \hat{\rho} \) with \( \hat{\rho} \) being a spherical distribution \( \hat{\rho}(\xi) = \phi(\|\xi\|) \) for some function \( \phi \), t
\[ \mathbb{E}[\text{clip}(g + \xi, c)] = r \cdot g, \forall g \in \mathbb{R}^d, \]  
where \( r \) is some constant. To see this, consider two vectors \( \xi_1 \) and \( \xi_2 \) satisfying
\[ \|\xi_1\| = \|\xi_2\|, \cos(\xi_1, g) = \cos(\xi_2, g), \sin(\xi_1, g) = -\sin(\xi_2, g). \]  
Then it is easy to see that \( \|\xi_1 + g\| = \|\xi_2 + g\| \), and \( \xi_1 + \xi_2 \) aligns with \( g \). It follows that 
\[ \text{clip}(g + \xi_1, c) + \text{clip}(g + \xi_2, c) = \frac{g + \xi_1}{\|g + \xi_1\|} \cdot \min\{c, \|g + \xi_1\|\} + \frac{g + \xi_2}{\|g + \xi_2\|} \cdot \min\{c, \|g + \xi_2\|\} \]
\[ = g + \xi_1 + g + \xi_2 \cdot \min\{c, \|g + \xi_1\|\} = v \cdot g, \]
for some constant \( v \). That is, \( \text{clip}(g + \xi_1, c) + \text{clip}(g + \xi_2, c) \) aligns with \( g \). Now let \( \xi_1 = g + v \) and \( \xi_2 = g - v \), we can see that (39) will be satisfied. Then we can integrate such pairs over the spherical distribution \( \hat{\rho}(\xi) \) and obtain (38).

Thus, the expected clipped gradient is in the same direction as \( g \). Combining the above relation with restated Theorem 2 above, it follows that when \( \hat{\rho} \) is a spherical distribution with \( \hat{\rho}(\xi) = \phi(\|\xi\|) \),
\[ \mathbb{E}[\text{clip}(g + \xi, c)] = r \cdot g. \]  
with \( r \geq 0 \) and
\[ r \cdot \|g\| \geq \min\left(\frac{3}{4}c, \|g\|\right) \mathbb{P}(\|\xi\| < \frac{c}{4}). \]

Now we can use the above results to prove the theorem.

The expectation can be splitted as
\[ \mathbb{E}_{\xi, \sim p}[(\nabla f(x_i), g_t)] = \sum_{i=1}^{m} w_i \mathbb{E}_{\xi, \sim p, i}[(\nabla f(x_i), g_t)]. \]

Then, because (40) and \( \mathbb{E}_{\xi, \sim p, i} [g_t] = u_t \) and that \( p_i \) corresponds to a noise with spherical distribution added to \( u_t \), we have
\[ \mathbb{E}_{\xi, \sim p}[(\nabla f(x_i), g_t)] = \langle \nabla f(x_i), \mathbb{E}_{\xi, \sim p, i} [g_t] \rangle = \langle \nabla f(x_i), r_i u_t \rangle, \]
with \( r_i \|u_t\| \geq \min\left(\frac{3c}{4}, \|u_t\|\right) \mathbb{P}_{\xi, \sim p, i}(\|\xi\| < \frac{c}{4}) \). Since we assumed \( \langle u_t, \nabla f(x_i) \rangle \geq 0 \), we have
\[ \mathbb{E}_{\xi, \sim p}[(\nabla f(x_i), g_t)] \geq \|\nabla f(x_i)\| \sum_{i=1}^{m} w_i \min\left(\frac{3c}{4}, \|u_t\|\right) \cos(u_t, \nabla f(x_i)) \mathbb{P}_{\xi, \sim p, i}(\|\xi\| < \frac{c}{4}) \geq 0. \]
This completes the proof. \( \square \)
where we define $g_{t,i} \triangleq \nabla f(x_t) + \xi_{t,i}$ is the stochastic gradient at iteration $t$ evaluated on sample $i$, and $S_t$ is a subset of whole dataset $D$; $Z_t \sim \mathcal{N}(0, \sigma^2 I)$ is the noise added for privacy. We denote $g_t := \frac{1}{|S_t|} \sum_{i \in S_t} \text{clip}(\nabla f(x_t) + \xi_{t,i}, c)$ in the remaining parts of the proof to simplify notation. It is clear that $\|g_t\| \leq c$.

Following traditional convergence analysis of SGD using smoothness assumption, we first have:

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{G}{2} \|x_{t+1} - x_t\|^2 \tag{41}$$

Taking expectation conditioned on $x_t$, we have

$$\mathbb{E}[f(x_{t+1})|x_t] \leq f(x_t) - \alpha \mathbb{E}[\langle \nabla f(x_t), g_t \rangle | x_t] + \frac{G \alpha^2}{2} \mathbb{E}[\|g_t\|^2 | x_t] + \sigma^2 d) \tag{42}$$

Take overall expectation and sum over $t \in [T]$ and rearrange, we have

$$\sum_{t=1}^{T} \alpha \mathbb{E}[\langle \nabla f(x_t), g_t \rangle \leq f(x_1) - \mathbb{E}[f(x_{T+1})] + \frac{T}{2} G \alpha^2 (c^2 + \sigma^2 d). \tag{43}$$

Dividing both sides by $T \alpha$, we get

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\langle \nabla f(x_t), g_t \rangle \leq f(x_1) - \mathbb{E}[f(x_{T+1})] + \frac{1}{2} G \alpha (c^2 + \sigma^2 d). \tag{44}$$

To achieve $(\epsilon, \delta)$-privacy, we need $\sigma^2 = \nu \frac{T \epsilon^2 \ln(\frac{1}{\delta})}{n^2 \epsilon^2}$ for some constant $\nu$ by Theorem 1 in [Abadi et al. 2016b]. Substituting the expression of $\sigma^2$ into the above inequality, we get

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\langle \nabla f(x_t), g_t \rangle \leq \frac{D_f}{T \alpha} + \frac{1}{2} G \alpha \left( c^2 + \nu \frac{T \ln(\frac{1}{\delta})}{n^2 \epsilon^2} c^2 d \right), \tag{45}$$

where we define $D_f := f(x_1) - \min_x f(x)$.

Setting $T \alpha = \sqrt{\frac{D_f}{n \epsilon \sqrt{\ln(\frac{1}{\delta})}}}$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\langle \nabla f(x_t), g_t \rangle \leq \left( \frac{1}{2} \nu + 1 \right) \frac{\sqrt{D_f G d \ln(\frac{1}{\delta})}}{n \epsilon} + \frac{1}{2} G \alpha \epsilon^2. \tag{47}$$
with \( \tilde{\rho} \) being a symmetric distribution. Applying Theorem 2, we have
\[
\mathbb{E}[\langle \nabla f(x_t), g_t \rangle] = \mathbb{E}_{\xi_t \sim p}[\langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle],
\]
with \( \xi_t \) being a discrete random variable that can takes values \( \xi_{i,t}, i \in D \) with equal probability and \( D \) is the whole dataset.

Now it is time to split the bias as following.
\[
\mathbb{E}_{\xi_t \sim \rho}[\langle \nabla f(x_t), g_t \rangle] = \mathbb{E}_{\xi_t \sim \rho}[\langle \nabla f(x_t), g_t \rangle] + \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle (p_t(\xi_t) - \tilde{p}_t(\xi_t)) d\xi_t,
\]
with \( \tilde{\rho} \) being a symmetric distribution. Applying Theorem 2, we have
\[
\mathbb{E}_{\xi_t \sim \rho}[\langle \nabla f(x_t), g_t \rangle] \geq \| \nabla f(x_t) \|^2 \cdot \mathbb{P}_{\xi_t \sim \rho}([\| \xi_t \| < \frac{c}{4}], \text{ if } \| \nabla f(x_t) \| \leq \frac{3}{4} c
\]
\[
\mathbb{E}_{\xi_t \sim \rho}[\langle \nabla f(x_t), g_t \rangle] \geq \frac{3}{4} c \cdot \mathbb{P}_{\xi_t \sim \rho}([\| \xi_t \| < \frac{c}{4}], \text{ if } \| \nabla f(x_t) \| \geq \frac{3}{4} c.
\]

Now we bound the bias term using the Wasserstein distance as follows.
\[
- \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle (p_t(\xi_t) - \tilde{p}_t(\xi_t)) d\xi_t
\]
\[
= \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle (\tilde{\rho}(\xi_t) - p(\xi_t)) d\xi_t
\]
\[
= \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle (\tilde{\rho}(\xi_t) d\xi_t - \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle (p(\xi_t) d\xi_t)
\]
\[
= \int \int |\langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle - \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t', c) \rangle | \gamma(\xi_t, \xi_t') d\xi_t d\xi_t'
\]
where \( \gamma \) is any joint distribution with marginal \( \tilde{\rho} \) and \( p \). Thus, we have
\[
- \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle (p_t(\xi_t) - \tilde{p}_t(\xi_t)) d\xi_t
\]
\[
\leq \inf_{\gamma \in \Gamma(\tilde{\rho}, p)} \int \int |\langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle - \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t', c) \rangle | \gamma(\xi_t, \xi_t') d\xi_t d\xi_t'
\]
where \( \Gamma(\tilde{\rho}, p) \) is the set of all coupling with marginals \( \tilde{\rho} \) and \( p \) on the two factors, respectively. Define the distance function \( d_{\gamma, c}(a, b) = |\langle y, \text{clip}(y + a, c) \rangle - \langle y, \text{clip}(y + b, c) \rangle| \), we have
\[
- \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + \xi_t, c) \rangle (p_t(\xi_t) - \tilde{p}_t(\xi_t)) d\xi_t
\]
\[
\leq \inf_{\gamma \in \Gamma(\tilde{\rho}, p)} \int \int d_{\gamma, c}(\xi_t, \xi_t') \gamma(\xi_t, \xi_t') d\xi_t d\xi_t' := W_{\gamma, c}(\tilde{\rho}, p),
\]
(52)
where $W_{v,c}(p,p')$ is the Wasserstein distance between $p$ and $p'$ using the metric $d_{v,c}$.

In summary, define

$$h(y) = \begin{cases} y^2, & \text{for } y \leq 3c/4 \\ \frac{3c}{4} y, & \text{for } y > 3c/4 \end{cases}$$

we have

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_{\xi_t \sim \tilde{p}}(\|\xi_t\| < \frac{c}{4}) h(\|\nabla f(x_t)\|) \leq \left(\frac{1}{2} v + \frac{3}{2}\right) c \sqrt{D_f Gd \ln \left(\frac{1}{\delta}\right)} \cdot \frac{1}{ne} + \frac{1}{T} \sum_{t=1}^{T} W_{v,c}(\tilde{p}_t, p_t). \quad (53)$$

This completes the proof. \hfill \square

### E Proof of Theorem 6

**Theorem 6** Let $g_t := \text{clip}(\nabla f(x_t) + \xi_t + k \zeta_t, c)$ and $\zeta_t \sim \mathcal{N}(0, I)$. Then gradient clipping algorithm has following properties:

$$\mathbb{E}_{\zeta_t \sim \mathcal{N}(0, I)} \{\nabla f(x_t), g_t\} \geq \|\nabla f(x_t)\| \min\left\{\|\nabla f(x_t)\|, \frac{3}{4} c\right\} \mathbb{P}(\|k \zeta_t\| < \frac{c}{4}) - \|\nabla f(x_t)\| O\left(\frac{\sigma_{\xi_t}^2}{k^2}\right), \quad (54)$$

where $\sigma_{\xi_t}^2$ is the variance of the gradient noise $\xi_t$.

**Proof**: Define $W_t = \xi_t + k \zeta_t$ as the total noise added to the gradient $\nabla f(x_t)$ before clipping; Also let us denote $W_t \sim \tilde{p}$. We know the following about $W_t$:

$$\mathbb{E}[W_t] = 0, \quad \tilde{p}(W_t) = \int_{\xi_t} p(\xi_t) \frac{1}{k} \psi\left(\frac{W_t - \xi_t}{k}\right) d\xi_t, \quad \text{with } \psi \text{ being the pdf of } \mathcal{N}(0, I). \quad (55)$$

The proof idea is to bound the total variation distance between $\tilde{p}(W_t)$ and $\frac{1}{k} \psi(\cdot)$ as $O\left(\frac{\sigma_{\xi_t}^2}{k^2}\right)$, then use this distance to bound the clipping bias $b_t$. This implies $\tilde{p}(W_t)$ will become more and more symmetric as $k$ increases.

Towards this end, we have

$$\int_{W_t} \left|\tilde{p}(W_t) - \frac{1}{k} \psi\left(\frac{W_t}{k}\right)\right| dW_t = \int_{W_t} \left|\int_{\xi_t} p(\xi_t) \frac{1}{k} \psi\left(\frac{W_t - \xi_t}{k}\right) d\xi_t - \frac{1}{k} \psi\left(\frac{W_t}{k}\right)\right| dW_t = k \int_{W_t'} \left|\int_{\xi_t} p(\xi_t) \frac{1}{k} \psi\left(\frac{W_t' - \xi_t}{k}\right) d\xi_t - \frac{1}{k} \psi(W_t')\right| dW_t'. \quad (56)$$

By Taylor's series, we have

$$\psi\left(W_t' - \frac{\xi_t}{k}\right) = \psi(W_t') + \left\langle \nabla \psi(W_t'), -\frac{\xi_t}{k}\right\rangle + \int_0^1 \frac{\xi_t}{k} \nabla^2 \psi\left(W_t' - \tau \frac{\xi_t}{k}\right) \frac{\xi_t}{k} (1 - \tau) d\tau. \quad (57)$$
Plugging the above into (56), we obtain:

\[
\int \left| \tilde{p}(W_t) - \frac{1}{k} \psi\left( \frac{W_t}{k} \right) \right| dW_t \\
= \int_{W_t'} \left| \int_{\xi_t} p(\xi_t) \psi \left( \frac{W_t' - \xi_t}{k} \right) d\xi_t - \psi(W_t') \right| dW_t'
\]

\[
= \int_{W_t'} \left| \int_{\xi_t} p(\xi_t) \int_{0}^{1} \left( \frac{\xi_t}{k}, \nabla^2 \psi(W_t' - t \frac{\xi_t}{k}) \right) (1 - \tau) d\tau d\xi_t \right| dW_t'
\]

\[
\leq \int_{0}^{1} \int_{\xi_t} \left| p(\xi_t) \left( \frac{\xi_t}{k}, \nabla^2 \psi(W_t' - t \frac{\xi_t}{k}) \right) (1 - \tau) \right| dW_t' d\xi_t d\tau,
\]

(58)

where the second equality is obtained by applying (57) and using the fact that \( \xi_t \) is zero mean. Noticing that \( \tau \leq 1 \) and define \( \hat{W}_t = W_t' - \tau \frac{\xi_t}{k} \), we have

\[
\int_{W_t'} \left| p(\xi_t) \left( \frac{\xi_t}{k}, \nabla^2 \psi(W_t' - \tau \frac{\xi_t}{k}) \right) (1 - \tau) \right| dW_t'
\]

\[
=p(\xi_t)(1 - \tau) \int_{\hat{W}_t} \left| \left( \frac{\xi_t}{k}, \nabla^2 \psi(W_t' \frac{\xi_t}{k}) \right) \right| dW_t'
\]

\[
=p(\xi_t)(1 - \tau) \int_{\hat{W}_t} \left| \left( \frac{\xi_t}{k}, \nabla^2 \psi(\hat{W}_t) \frac{\xi_t}{k} \right) \right| d\hat{W}_t,
\]

\[
=p(\xi_t)(1 - \tau) \int_{R\hat{W}_t} \left| \left( \frac{R\xi_t}{k}, \nabla^2 \psi(R\hat{W}_t) \frac{R\xi_t}{k} \right) \right| dR\hat{W}_t
\]

\[
=p(\xi_t)(1 - \tau) \int_{\hat{W}_t} \left| \left( \frac{R\xi_t}{k}, \nabla^2 \psi(\hat{W}_t) \frac{R\xi_t}{k} \right) \right| d\hat{W}_t,
\]

(59)

where \( R \) is an arbitrary rotation matrix which means the integration term only depends on \( \| \frac{\xi_t}{k} \| \). Thus we can assume \( \xi_{t,1} = \| \xi_t \| \) and \( \xi_{t,i} = 0 \) for \( i \geq 2 \), wlog. Then, we have

\[
\int_{W_t'} \left| p(\xi_t) \left( \frac{\xi_t}{k}, \nabla^2 \psi(W_t' - \tau \frac{\xi_t}{k}) \right) (1 - \tau) \right| dW_t'
\]

\[
\leq p(\xi_t)(1 - \tau) \int_{W_t'} \left( \frac{\xi_t}{k} \right)^2 \left| \nabla^2_{1,1} \psi(W_t' - \tau \frac{\xi_t}{k}) \right| dW_t'
\]

\[
\leq p(\xi_t)(1 - \tau) \int_{\hat{W}_t} \left( \frac{\xi_t}{k} \right)^2 \left| \nabla^2_{1,1} \psi(\hat{W}_t) \right| |Det\left( \frac{dW_t'}{d\hat{W}_t} \right)| d\hat{W}_t
\]

\[
\leq p(\xi_t)(1 - \tau) \left( \frac{\| \xi_t \|}{k} \right)^2 q,
\]

(60)

where we have define \( \hat{W}_t = W_t' - \tau \frac{\xi_t}{k} \) and \( q = \int_{-\infty}^{\infty} |h''(x)| dx \) with \( h(x) \) being the pdf of 1-dimensional standard normal distribution. Thus, \( q \) is a dimension independent constant.
Substituting (60) into (58), we get
\[
\int \left| \bar{p}(W_t) - \frac{1}{k} \psi \left( \frac{W_t}{k} \right) \right| dW_t \\
\leq \int_0^1 \int_{\xi_t}^{W_t} p(\xi_t) \left\langle \frac{\xi_t}{k}, \nabla^2 \psi(W'_t - \tau \frac{\xi_t}{k}) \right\rangle (1 - \tau) \left\| \xi_t \right\|^2 q d\xi_t d\tau \\
\leq \int_0^1 \int_{\xi_t}^{W_t} p(\xi_t) (1 - \tau) \left\| \xi_t \right\|^2 q d\xi_t d\tau \\
= \int_0^1 (1 - \tau) \frac{\sigma_{\xi_t}^2}{k^2} q d\tau \\
= \frac{1}{2} \frac{\sigma_{\xi_t}^2}{k^2} q.
\] (61)

where we used the fact that \( \mathbb{E}[\xi_t] = 0 \) and defined \( \sigma_{\xi_t}^2 \) being the variance of \( \xi_t \).

By (5), we know that the following holds:

\[
\mathbb{E}_{\xi_t \sim p, c_t} [\langle \nabla f(x_t), g_t \rangle] = \mathbb{E}_{W_t \sim \tilde{p}} [\langle \nabla f(x_t), g_t \rangle] \\
+ \int \langle \nabla f(x_t), \text{clip}(\nabla f(x_t) + W_t, c) \rangle (p_t(W_t) - \tilde{p}_t(W_t)) dW_t.
\] (62)

Let \( \tilde{p} \) be the pdf of \( k \xi_t \), from Theorem 2 we have

\[
\mathbb{E}_{W_t \sim \tilde{p}} [\langle \nabla f(x_t), g_t \rangle] \geq \| \nabla f(x_t) \| \min \left\{ \frac{3}{4} c, \| \nabla f(x_t) \| \right\} \mathbb{P}(\| k \xi_t \| \leq \frac{c}{4}).
\] (63)

In addition, we can bound \( b_t \) as

\[
|b_t| \leq \| \nabla f(x_t) \| \cdot c \cdot \int |p_t(\xi_t) - \tilde{p}_t(\xi_t)| d\xi_t \\
\stackrel{(61)}{\leq} \| \nabla f(x_t) \| \cdot \frac{c}{2} \cdot \frac{\sigma_{\xi_t}^2}{k^2} q = \| \nabla f(x_t) \| O \left( \frac{\sigma_{\xi_t}^2}{k^2} \right).
\] (64)

Combining (62), (64), and (63) finishes the proof.

\[ \Box \]

\section*{F More experiments on random projection}

In this section, we show the projection of stochastic gradients into 2d spaces described in Section 4 for different projection matrices in Figure 4. It can be seen that as the training progresses, the gradient distribution in 2d space tends to be increasingly more symmetric.
Figure 4: Distribution of gradients on MNIST after epochs 0 projected using different random matrices.

Figure 5: Distribution of gradients on MNIST after epochs 3 projected using different random matrices.

Figure 6: Distribution of gradients on MNIST after epochs 9 projected using different random matrices.

27
G Evaluation on the probability term

In this section, we evaluate the probability term in Corollary 1 using a few statistics of the empirical gradient distribution on MNIST. Specifically, at the end of different epochs, we plot histogram of norm of stochastic gradient and norm of noise, along with the inner product between stochastic gradient (and clipped stochastic gradient) and the true gradient. The results are shown in Figure 7. One observation is that the norm of stochastic gradients is concentrated around 0 while having a heavy tail. The noise distribution is concentrated around some positive value with a heavy tail, the mode of the noise actually corresponds to the approximate 0 norm mode of stochastic gradients. As the training progresses, the norm of stochastic gradients and the norm of noise are approaching 0. We set clipping threshold to be 1 in the experiment, so actually the probability \( P(\|\xi_t\| \leq T/4) \) is 0 for the empirical distribution \( p \). When we use a distribution \( \tilde{p} \) with \( P(\|\xi_t\| \leq T/4) \geq l \) for some value \( l > 0 \) to approximate \( p \), this approximation indeed can create an approximation bias. However, the bias may not be too large since the mode of the norm of noise is not too much bigger than \( T/4 \). Furthermore, in Corollary 1 and Theorem 2, we actually can change \( P_{\xi_t \sim p}(\|\xi_t\| \leq T/4) \) to \( P_{\xi_t \sim \tilde{p}}(\|\xi_t\| \leq zc) \) with any \( z < 1 \) and simultaneously change the \( T/4 \) to \((1 - z)c\) to make the probability term larger.

Despite the discussions above, the distribution of the norm of stochastic gradients and the norm of the noise combined with the 2d visualization experiments implies the noise on gradient might follow a mixture of distributions with each component being approximate symmetric. Specifically, one component may correspond to an approximate 0 mean distribution of stochastic gradients. Intuitively this can be true since each class of data may corresponds to a few variations of stochastic gradients and the gradient noise is observed to be low rank in Li et al. [2020]. We have some discussions in Section 2.2 to explain how convergence can be achieved in the cases of symmetric distribution mixtures but it may not be the complete explanation here. Further exploration of gradient distribution in practice is an important question and we leave it for future research.
Figure 8: Distribution of different statistics at epoch 3.

Figure 9: Distribution of different statistics at epoch 9.
**Figure 10:** Distribution of different statistics at epoch 59.

**Table 1: Scalability of** $E_{\xi_t=0,\zeta_t}[\langle \nabla f(x_t), g_t \rangle]$ **w.r.t.** $d$ and $k$

| $k$  | $d = 1$ | $d = 10$ | $d = 100$ | $d = 1,000$ | $d = 10,000$ |
|------|---------|-----------|------------|-------------|--------------|
| 1    | 10      | 9.572     | 7.077      | 3.015       | 0.995        |
| 10   | 6.788   | 2.961     | 0.992      | 0.316       | 0.1          |
| 100  | 0.758   | 0.316     | 0.098      | 0.032       | 0.01         |
| 1000 | 0.084   | 0.019     | 0.011      | 0.003       | 0.001        |

### Additional results and discussions on the probability term gradient correction in Section [5](#)

Theorem 6 says that after adding the Gaussian noise $k\zeta_t$ before clipping, the clipping bias can decrease. Meanwhile, the expected descent also decreases because $P(||k\zeta_t|| < \frac{c}{4})$ decreases with $k$. To get a more clear understanding of the theorem, consider $d = 1$, then $P(||k\zeta_t|| < \frac{c}{4}) = \text{erf}(\frac{c}{4k})$ which decreases with an order of $O\left(\frac{1}{k}\right)$. This rate is slower than the $O\left(\frac{1}{k^2}\right)$ diminishing rate of the clipping bias. Thus, as $k$ becomes large, the clipping bias will be negligible compared with the expected descent. This will translate to a slower convergence rate with a better final gradient bound in convergence analysis. The key idea of adding $k\zeta_t$ before clipping is to “symmetrify” the overall gradient noise distribution. By adding the isotropic symmetric noise $k\zeta_t$, the distribution of the resulting gradient noise $W_t \triangleq \xi_t + k\zeta_t$ will become increasingly more symmetric as $k$ increases. In particular, the total variation distance between the distribution of $W_t$ and $k\zeta_t$ decreases at a rate of $O\left(\frac{1}{k^2}\right)$ which can be further used to bound the clipping bias. Then, one can apply Theorem 2 to lower bound $E_{\xi_t=0,\zeta_t}[\langle \nabla f(x_t), g_t \rangle]$ by letting $\tilde{p}$ be the distribution of $k\zeta_t$. We believe the lower bounds in Theorem 6 can be further improved when $d > 1$, notice that $P(||k\zeta_t|| < \frac{c}{4})$ tends to
decrease fast with $k$ when $d$ being large.

However, we observe $E_{\xi_t \sim p_{\xi_t}} \left[ \langle \nabla f(x_t), g_t \rangle \right]$ decreases with a rate of $O(1/d)$ and $O(1/k)$ in practice for fixed $\|\nabla f(x_t)\|$ and $\xi_t = 0$ (see Table 1 for $\|\nabla f(x_t)\| = 10$, the expectation $E_{\xi_t \sim 0, \xi_t} \left[ \langle \nabla f(x_t), g_t \rangle \right]$ is evaluated over $10^5$ samples of $\xi_t \sim \mathcal{N}(0, I)$). In addition, one can prove that the lower bounds in Theorem 2 are tight up to a constant when $d = 1$ or $\tilde{p}(\xi_t)$ is a distribution on a one dimensional subspace. This implies the lower bound can only be improved by using more properties of isotropic distributions like $\mathcal{N}(0, I)$ or resorting to a more general form of the lower bounds. We found this to be non-trivial and decide to leave it for future research.