Understanding Gradient Clipping in Private SGD: A Geometric Perspective

Xiangyi Chen[1], Zhiwei Steven Wu[2], Mingyi Hong[1]

[1]University of Minnesota - Twin Cities, [2]Carnegie Mellon University

Contribution

- A symmetricity-based distribution-aware analysis on clipping bias
- Theoretical and empirical studies of the effect of gradient clipping in DP-SGD
- A pre-clipping perturbation mechanism to reduce clipping bias in DP-SGD

Motivation

- Private SGD with gradient clipping [1] works well in practice and the clipping threshold is an important parameter to tune
- Intuitively, gradient clipping may make Private SGD fail to converge
- The effect of gradient clipping is not well-understood

Differentially-private SGD (DP-SGD) and gradient clipping

Update rule:

$Z_t \sim \mathcal{N}(0, \sigma^2 I)$, the noise to achieve privacy

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

A subset of training samples

Per-sample gradient

Gradient clipping:

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

SGD with gradient clipping

To better understand convergence, first consider **SGD with gradient clipping** (batch size=1)**:**

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

clipped gradient

Convergence of SGD with gradient clipping

An intermediate convergence result:

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

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

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

However, what can it say about convergence?

1. ***Convergence***: We have $\mathbb{E}[\langle \nabla f(x_t), g_t\rangle] = \|\nabla f(x_t)\|^2$ when clipping is always inactive (c is very large).
2. ***<u>Divergence</u>***: The above may not hold when clipping can be active (c is relatively small).

Symmetricity helps convergence in SGD with gradient clipping

When distribution of stochastic gradient is symmetric:

Theorem. *Assume $\tilde{p}(\xi_t) = \tilde{p}(-\xi_t)$, gradient clipping with threshold c has the following properties.*

1. *If $\|\nabla f(x_t)\| \le \frac{3}{4}c$, then* $\mathbb{E}_{\xi_t\sim\tilde{p}}[\langle \nabla f(x_t), g_t\rangle] \ge \|\nabla f(x_t)\|^2 \mathbb{P}_{\xi_t\sim\tilde{p}}\left(\|\xi_t\| < \frac{c}{4}\right)$
2. *If $\|\nabla f(x_t)\| > \frac{3}{4}c$, then* $\mathbb{E}_{\xi_t\sim\tilde{p}}[\langle \nabla f(x_t), g_t\rangle] \ge \frac{3}{4}c\|\nabla f(x_t)\| \mathbb{P}_{\xi_t\sim\tilde{p}}\left(\|\xi_t\| < \frac{c}{4}\right)$

Implies convergence

Convergence of SGD and DP-SGD with gradient clipping

SGD with gradient clipping:

Theorem. *For SGD with gradient clipping, set $\alpha = \frac{1}{\sqrt{T}}$. Suppose true gradient noise distribution is p, choose $\tilde{p}_t(\xi) = \tilde{p}_t(-\xi)$, then the following holds:*

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

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

Clipping bias (Symmetricity-based)

DP-SGD with gradient clipping:

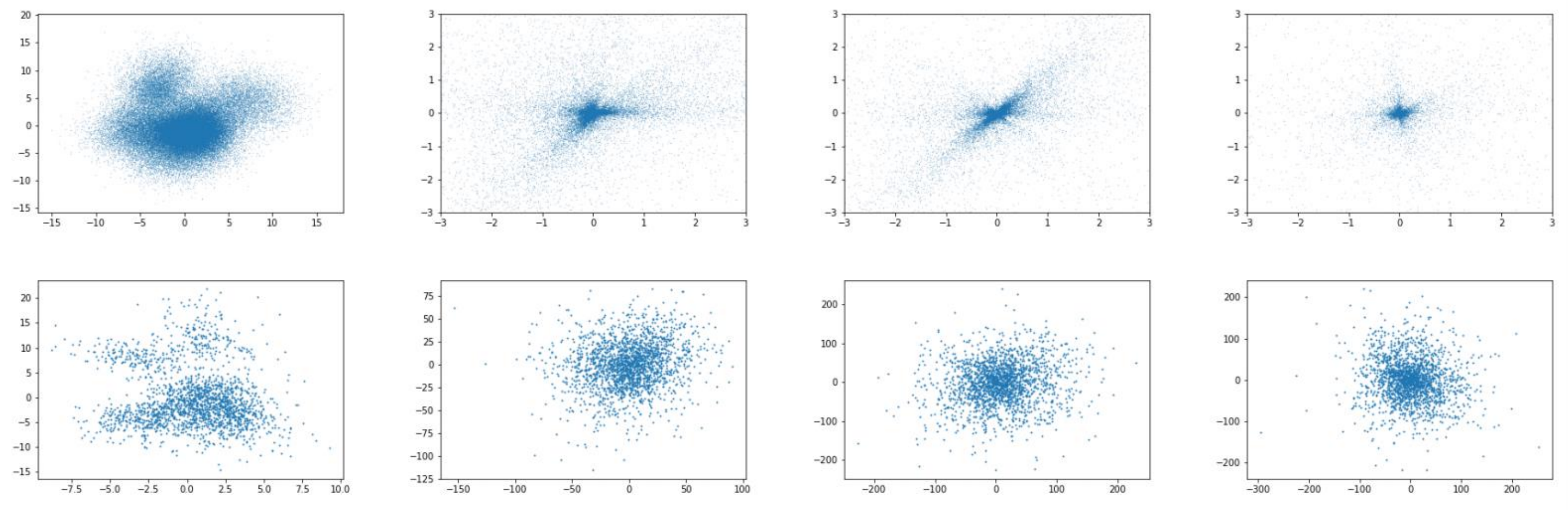
Theorem. *Let d be the dimensionality of the parameters. For DP-SGD with gradient clipping and privacy parameters (ϵ, δ), choose $\tilde{p}_t(\xi_t) = \tilde{p}_t(-\xi_t)$, there exist u and v such that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}_{\xi_t\sim\tilde{p}}\left(\|\xi_t\| < \frac{c}{4}\right)h_c(\|\nabla f(x_t)\|) \le \left(\frac{1}{2}v + \frac{3}{2}\right)c\frac{\sqrt{D_f G d\ln(\frac{1}{\delta})}}{n\epsilon} + \frac{1}{T}\sum_{t=1}^{T} W_{\nabla f(x_t),c}(\tilde{p}_t, p_t)$$

where $h_c(y) = \min(y^2, \frac{3}{4}cy)$ and $W_{v,c}(p, p')$ is the Wasserstein distance between p and p' with metric function $d_{v,c}(a, b) = |\langle v, clip(v + a, c)\rangle - \langle v, clip(v + b, c)\rangle|$ and $D_f \ge f(x_1) - \min_x f(x)$.

Gradient symmetricity

Per-sample gradient of a convnet projected into 2d space using a random matrix (top row for MNIST, bottom row for CIFAR-10):



(a) Epoch 0 (b) Epoch 5 (c) Epoch 10 (d) Epoch 59

Mitigate clipping bias by pre-clipping noise

DP-SGD with pre-clipping noise:

$\zeta_{t,i} \sim \mathcal{N}(0, I)$

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

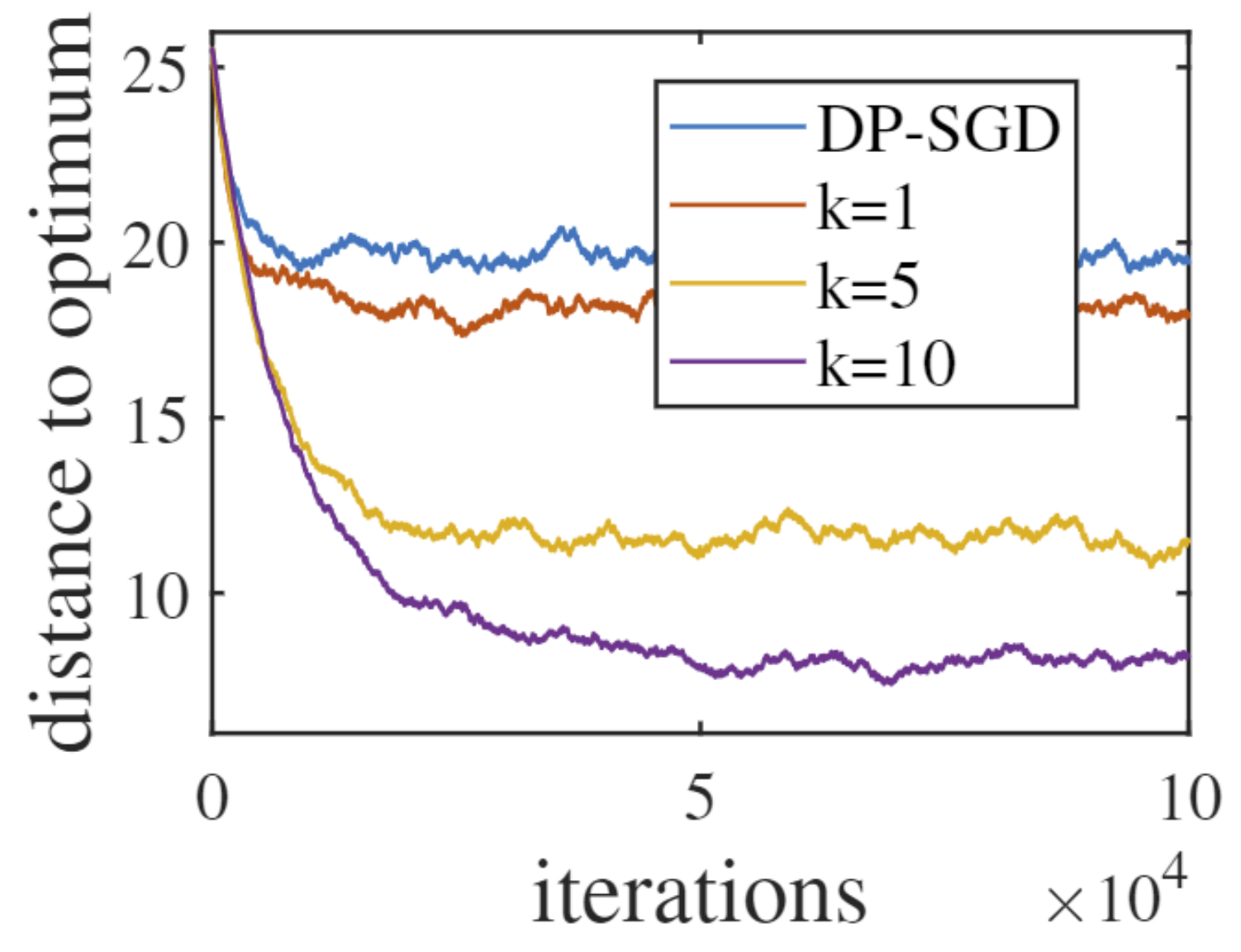
Theorem. *Let $g_t = clip(\nabla f(x_t) + \xi_t + k\zeta_t, c)$ and $\zeta_t \sim \mathcal{N}(0, I)$. Then gradient clipping algorithm has following properties:*

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

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

Benefit of pre-clipping noise

Run DP-SGD with pre-clipping noise
on a 10d synthetic dataset
with asymmetric gradient distribution



References

[1]. Abadi, Martin, et al. "Deep learning with differential privacy." 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016.