Lecture 14:
Differentially Private Deep Learning

Steven Wu
Carnegie Mellon University

Model Training with DP

Given private data $s_1, \ldots, s_n$, solve

$$\min_{x \in \mathbb{R}^d} f(x) \equiv \frac{1}{n} \sum_{i=1}^{n} \ell(x; s_i)$$

subject to differential privacy

DP-SGD (in Theory)

Differentially Private SGD [BST14, SCS13]

- At each iteration t,
  - Gradient estimate on a mini-batch S_t :
    $$g_t = \left(\frac{1}{|S_t|} \sum_{i \in S_t} \nabla \ell(x_t; s_i) \right)$$
  - Noisy gradient update :
    $$x_{t+1} = x_t - \eta \left(g_t + Z_t \right),$$
    $$Z_t \sim \mathcal{N}(0, \sigma^2 I_d)$$

Privacy Proof
assumes ℓ is L-Lipschitz
for all x

$\downarrow$

$\|\nabla \ell(x_t; s_i)\|_2 \leq L$

Set σ to scale with L

DP-SGD (in Practice)

Differentially Private SGD [ACGMMTZ16]

- At each iteration t,
  - For each s_i in the mini-batch S_t :

$$g_t = \left(\frac{1}{|S_t|} \sum_{i \in S_t} \mathrm{Clip}(\nabla \ell(x_t; s_i), C) \right)$$

  - Noisy gradient update :

$$x_{t+1} = x_t - \eta \left(g_t + Z_t\right),\ Z_t \sim \mathcal{N}(0, \sigma^2 I_d)$$

Gradient Clipping:

$$\mathrm{Clip}(g, C) = g \min \left\{ 1, \frac{C}{\|g\|_2} \right\}$$

Set σ to scale with C

Privacy Guarantee for DP-SGD (with Clipping)

[BST14, ACGMMTZ16]

Theorem: DP-SGD with gradient clipping of threshold C satisfies (ϵ, δ)-differential privacy, if the noise rate

$$\sigma \geq a \frac{C\, q \sqrt{T \ln(1/\delta)}}{\epsilon}$$

for some constant a and $q = \frac{|S_t|}{n}$.

How about convergence and optimality?

Part 1: Understanding the effects of clipping on differentially private optimization

Clipping is a special case of projection:

$\Pi_G(g) = \min_{g'} \|g - g'\|_2$ where G is ball of radius C

Projection onto G of other geometric structures

Part 2: Leveraging the low-dimensional structure in gradients to achieve better accuracy

Part 1: Understanding the effects of gradient clipping on private optimization

- *Xiangyi Chen, Z. S.W., Mingyi Hong*
  "Understanding Gradient Clipping in Private SGD: A Geometric Perspective"
  In NeurIPS 2020 (Spotlight)

Bad Example 1

Loss: $f(x) = \frac{1}{3}\sum_{i=1}^{3}\frac{1}{2}(x - s_i)^2$

where $s_1 = s_2 = -3$ and $s_3 = 9$.

$\Rightarrow$ Optimum $x^\star = 1$

Clipped gradient at $x^\star$

$\mathbb{E}[\text{Clip}(\nabla_x \ell(x^\star; s_i), 1)] = 0$

$\Rightarrow$ push iterates away from opt

Bad Example 2

Loss: $f(x) = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{2}(x - s_i)^2$

where $s_1 = 3,\ s_2 = -3$

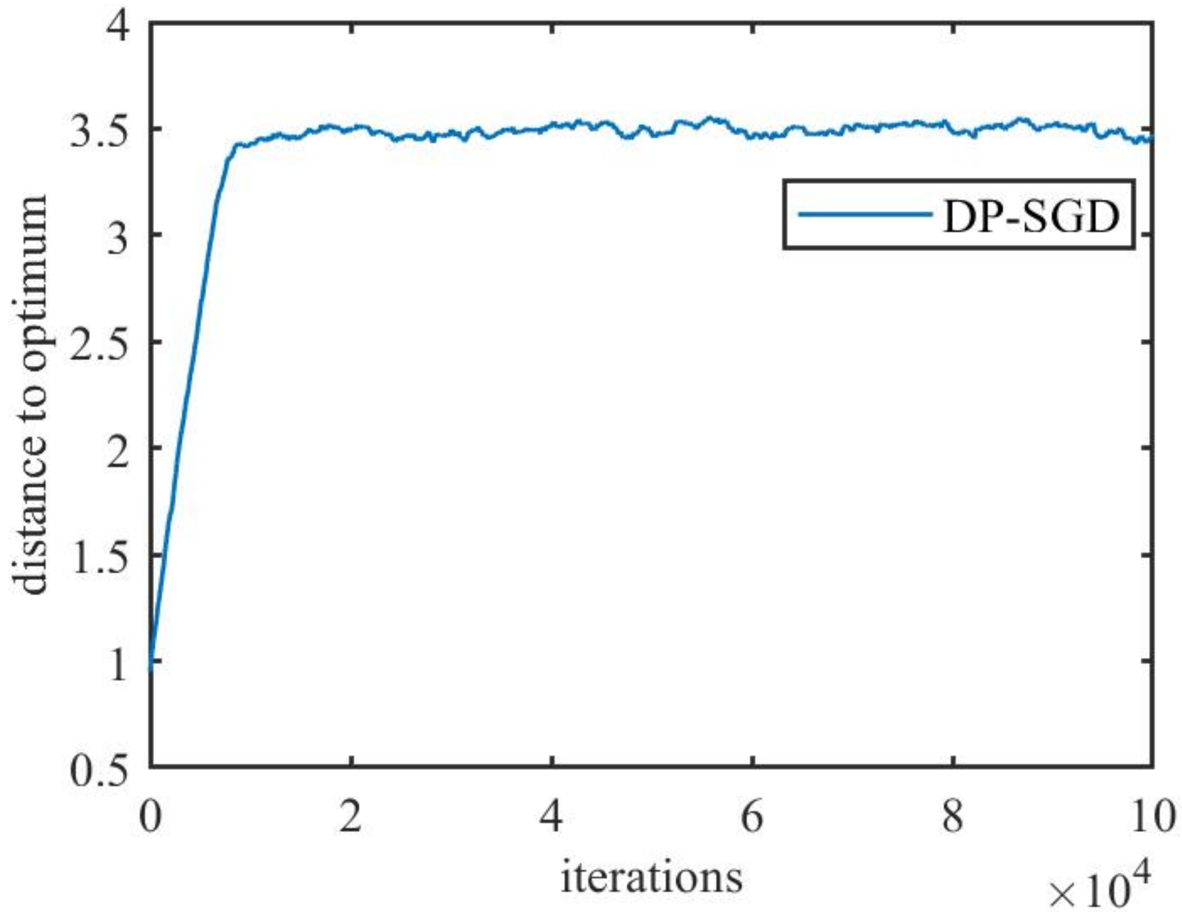
$\Rightarrow$ Optimum $x^\star = 0$

Clipped gradient for any $x \in [-2,2]$

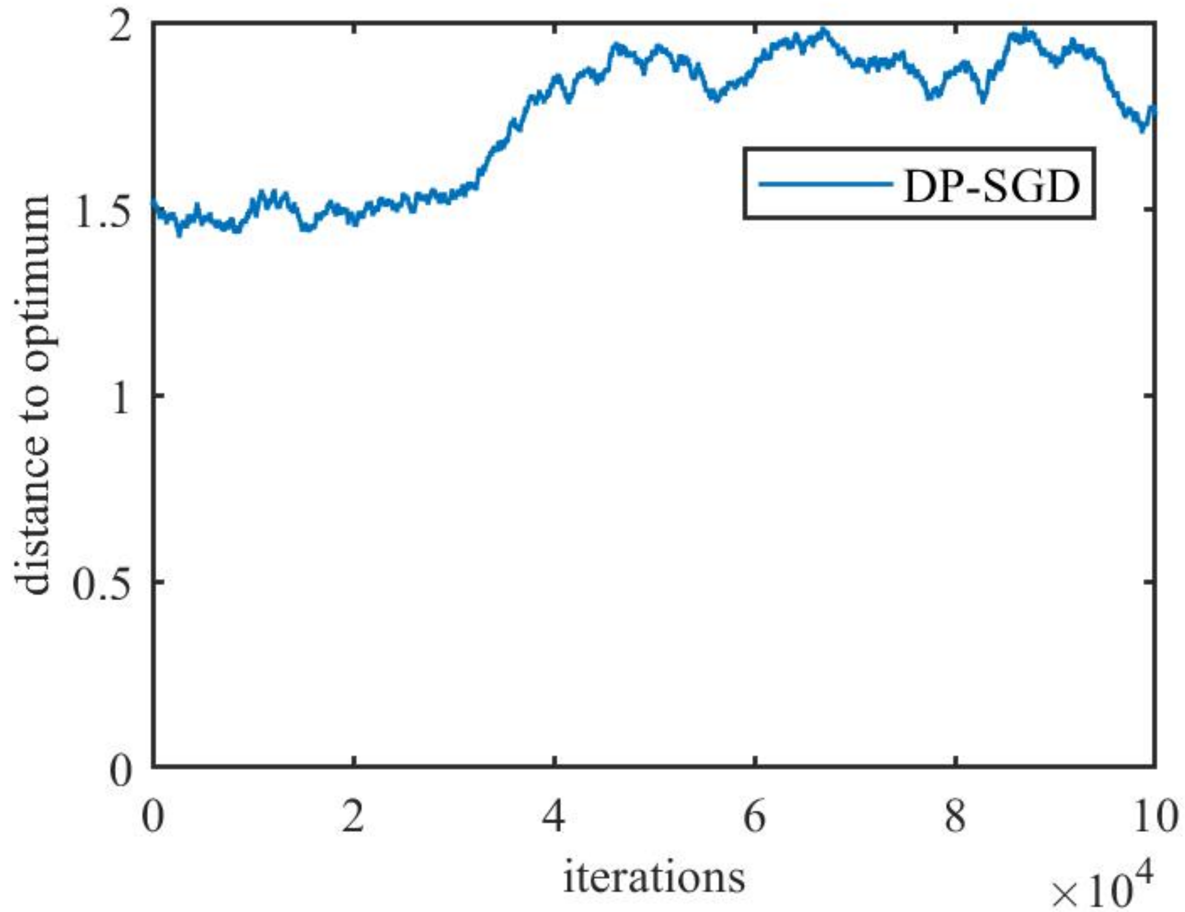
$$\mathbb{E}[\text{Clip}(\nabla_x \ell(x^\star; s_i), 1)] = 0$$

$\Rightarrow$ does not converge to opt

Adversarial Effects of Clipping



Example 1



Example 2

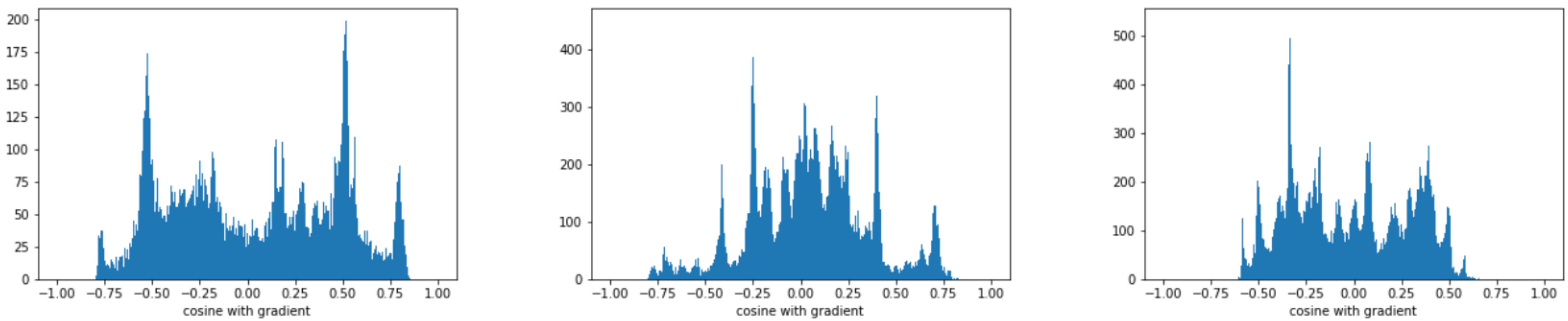
Do these occur in practical instances?

DP-SGD on MNIST

- DP-SGD with Clip norm $C = 1$
  60 epochs, $\epsilon \approx 3$, test accuracy $\approx 96.5\,\%$

- DP-SGD with Clip norm $C = 0.1$
  60 epochs, $\epsilon \approx 3$, test accuracy $\approx 92\,\%$

A glimpse of gradient distribution



(a) Epoch 4 (b) Epoch 10 (c) Epoch 59

Histogram of cosine between
stochastic gradients and true gradient

Does symmetricity in gradient distribution lead to convergence?

Analysis without DP

SGD with gradient clipping:

- Clipped gradient: $g_t = \text{Clip}\left(\nabla f(x_t) + \xi_t, C\right)$
  ξ_t : stochastic gradient noise

- $x_{t+1} = x_t - \eta\, g_t$

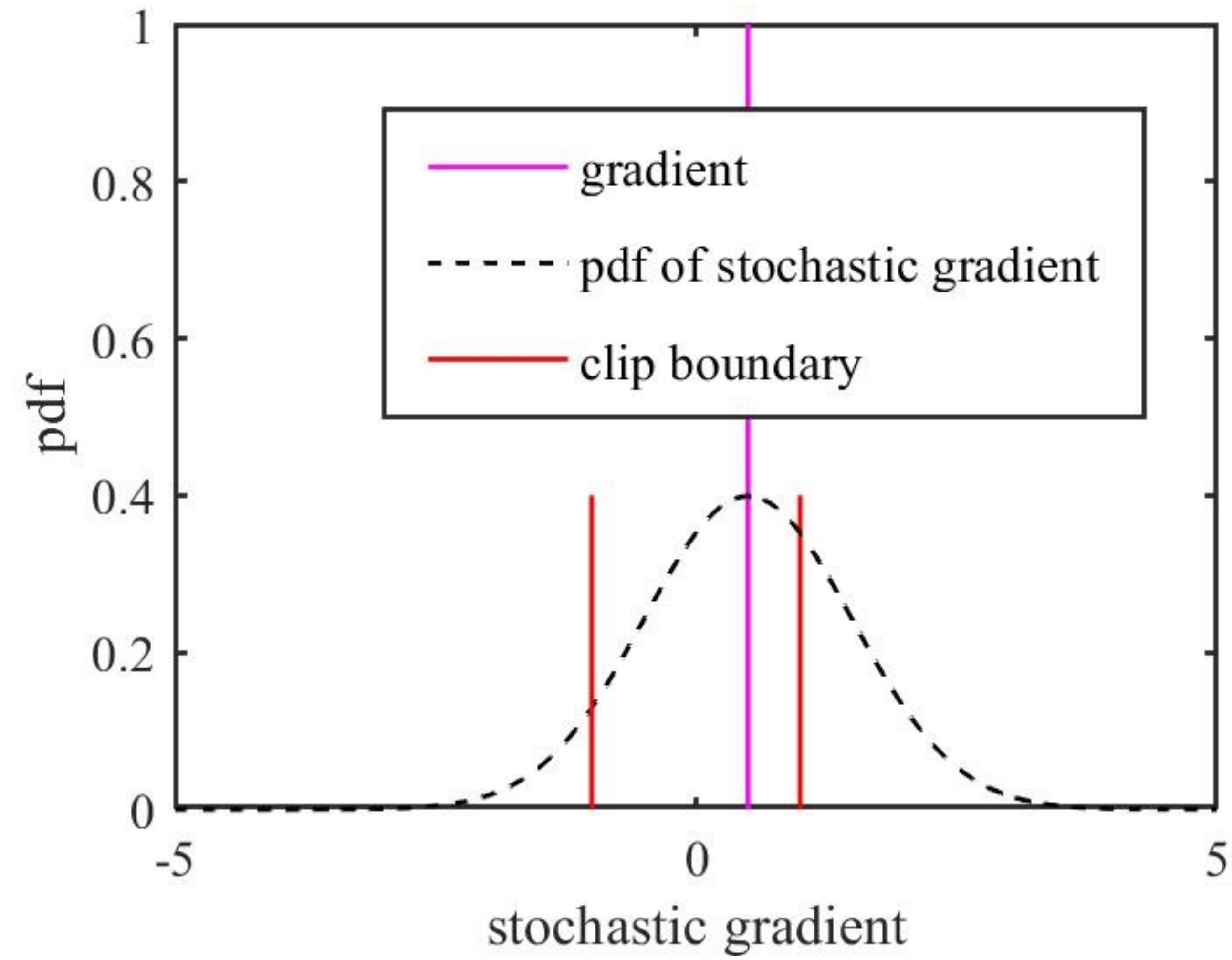
$$\boxed{\begin{array}{l}\text{Theorem [CWH20]. For } \eta = 1/\sqrt{T}\,, \\ \displaystyle \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\langle \nabla f(x_t), g_t \rangle\right] \le O\left(\frac{C^2}{\sqrt{T}}\right)\end{array}}$$

Want: $\sum_t \mathbb{E}[\langle \nabla f(x_t), g_t \rangle] \ge \sum_t \mathbb{E}[\|\nabla f(x_t)\|^2] \longrightarrow$ Converges to First-order stationary point

1-d Analysis

$$f(x) = \frac{1}{2}\mathbb{E}[(x-a)^2]$$
$$a \sim \mathcal{N}(0,1)$$

For $x = 0.5$, calculate
$\mathbb{E}[\text{Clip}(\nabla f(x_t) + \xi, 1)]$

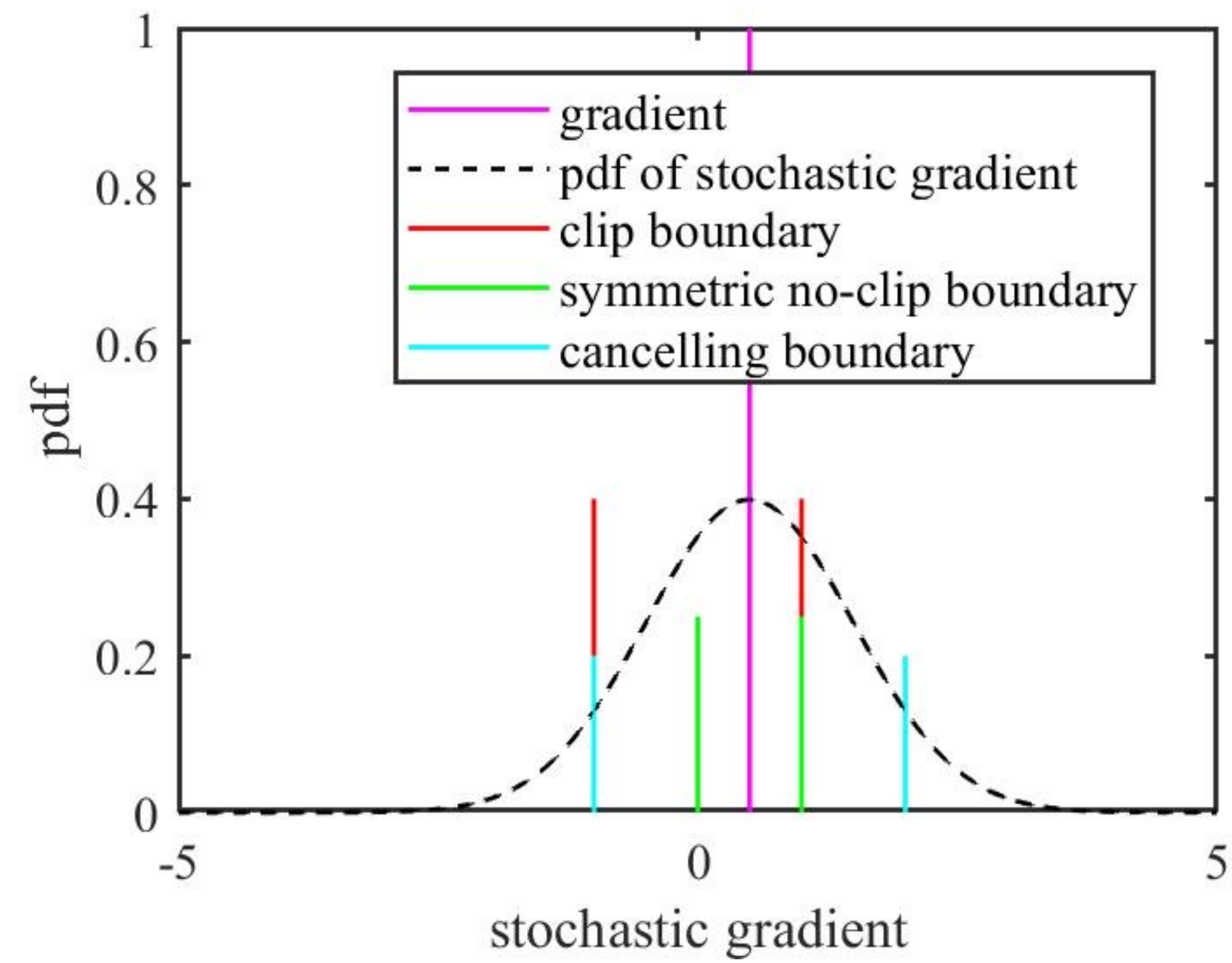


1-d Analysis

$$f(x) = \frac{1}{2}\mathbb{E}[(x-a)^2]$$
$$a \sim \mathcal{N}(0,1)$$

For $x = 0.5$, calculate $\mathbb{E}[\text{Clip}(\nabla f(x_t) + \xi, 1)]$

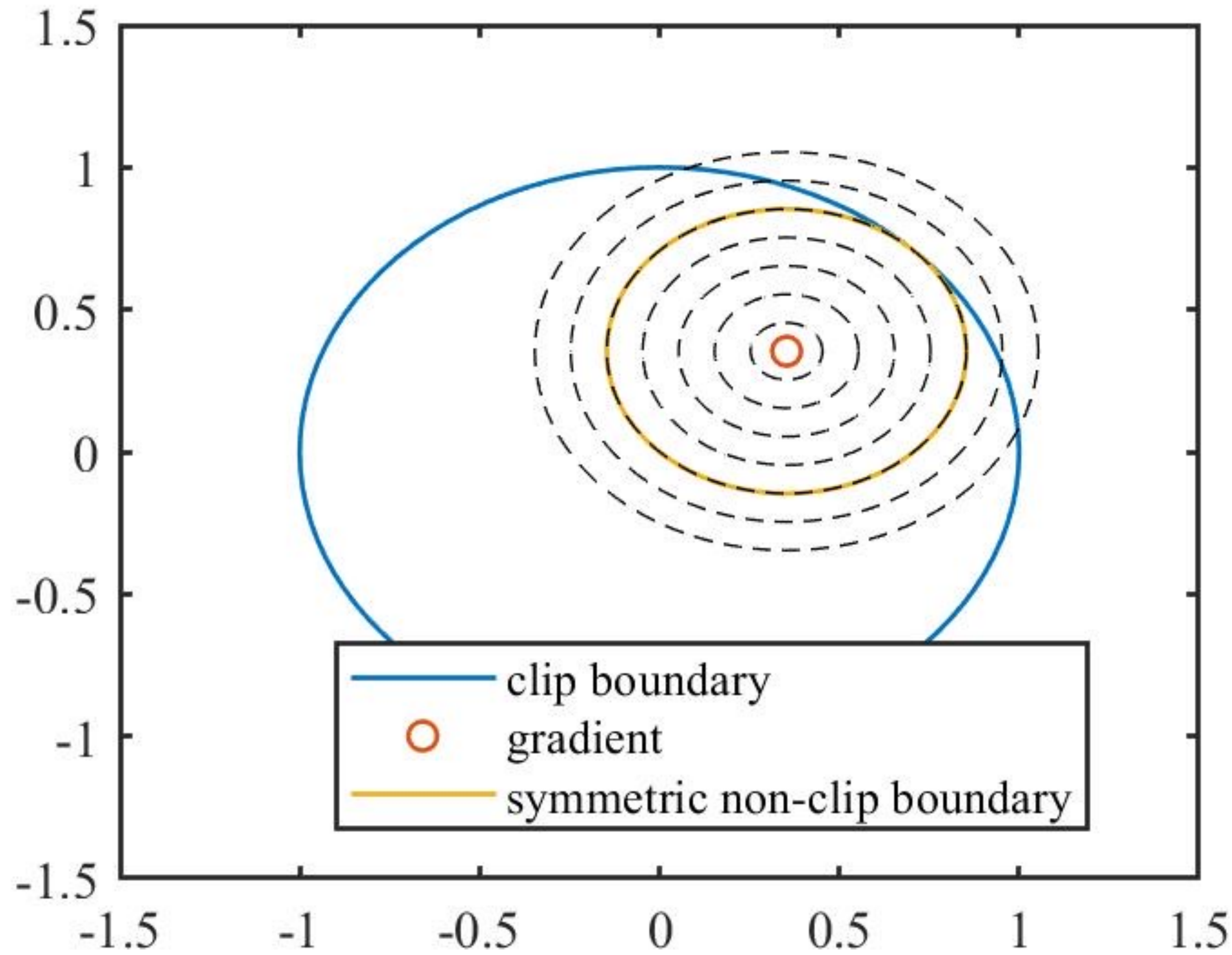
- Outside of cancelling boundary: stochastic gradient cancels
- Expectation = true gradient inside symmetric no-clip boundary



Higher Dimension

Symmetric gradient distribution

$$p(\xi) = p(-\xi)$$



Symmetricity-based Analysis

Clipped gradient: $g = \text{Clip}\left(\nabla f(x) + \xi, C\right)$

ξ : stochastic gradient noise

Theorem [CWH20]. Assume $\tilde{p}$ is a symmetric distribution: $\tilde{p}(\xi) = \tilde{p}(-\xi)$ for any $\xi \in \mathbb{R}^d$. Then clipped gradient g satisfies

$$\mathbb{E}_{\xi \sim \tilde{p}}\left[\langle \nabla f(x), g\rangle\right] \geq h\left(\|\nabla f(x)\|\right)\ \mathbb{P}_{\tilde{p}}\left[\|\xi\| < C/4\right]$$

where $h(y) = \min\left\{y^2, \dfrac{3Cy}{4}\right\}.$

Distributional Approximation

Couple gradient distribution p with a symmetric distribution $\tilde{p}$

$$\mathbb{E}_{\xi\sim p}\left[\langle \nabla f(x), g\rangle\right] = \mathbb{E}_{\xi\sim\tilde{p}}\left[\langle \nabla f(x), g\rangle\right] + \underbrace{\int \langle \nabla f(x), \mathrm{Clip}(\nabla f(x), C)\rangle \left(p(\xi) - \tilde{p}(\xi)\right)\, d\xi}_{\text{Clipping bias } b}$$

$\downarrow$

Bounded by Wasserstein distance $W_{\nabla f(x),c}(\tilde{p}, p)$

with metric $d_{v,C}(a,b) = |\langle v, \mathrm{Clip}(v+a, C)\rangle - \langle v, \mathrm{Clip}(v+b, C)\rangle|$

Convergence of SGD w/ Clipping

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

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

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{P}_{\xi_t \sim \tilde{p}_t}\left(\|\xi_t\| < \frac{c}{4}\right) \min\left\{\|\nabla f(x_t)\|, \frac{3}{4}c\right\} \|\nabla f(x_t)\| \leq \frac{D_f}{\sqrt{T}} + \frac{G}{2\sqrt{T}}c^2 - \frac{1}{T}\sum_{t=1}^{T} b_t, \tag{7}$$

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

Convergence of Clipped DP-SGD

Theorem [CWH20]. For each t, let $\tilde{p}_t$ be a "coupled" symmetric gradient distribution. Let $h(y) = \min\{y^2, \frac{3Cy}{4}\}$. Suppose that ∇f is Lipschitz, then

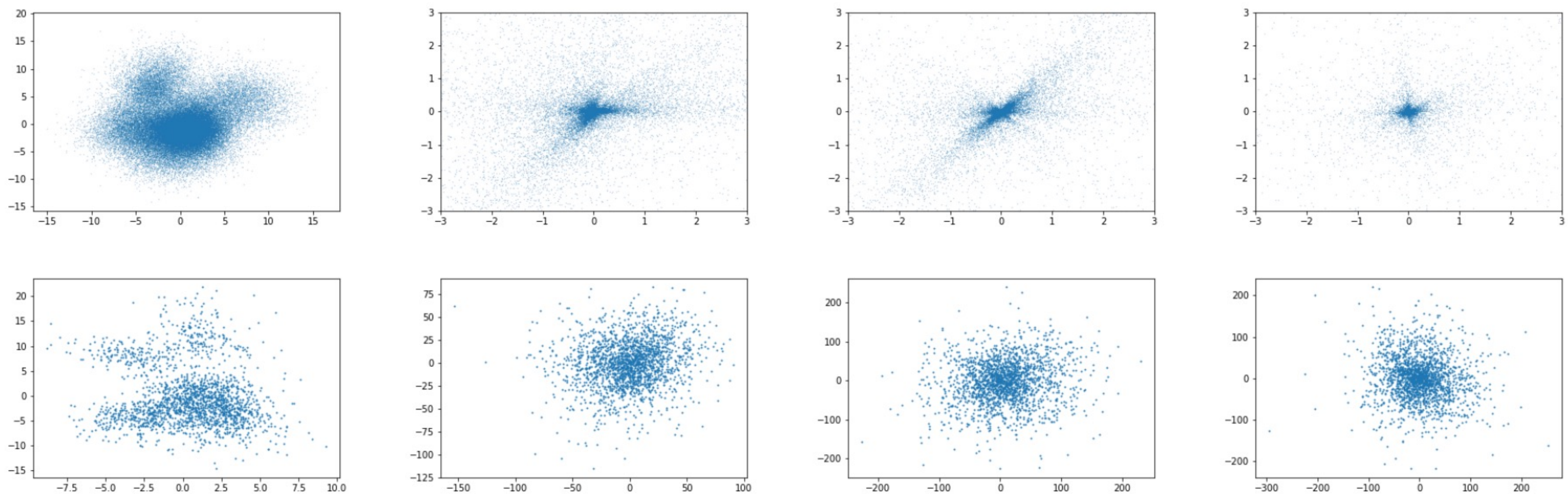
$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{P}_{\xi\sim\tilde{p}_t}\left[\|\xi\| < C/4\right]\ h(\|\nabla f(x_t)\|) \leq O\left(\frac{C\sqrt{d\ln(1/\delta)}}{n\epsilon}\right) + \frac{1}{T}\sum_{t=1}^{T} W_{\nabla f(x_t),C}(\tilde{p}_t, p_t)$$

DP-SGD without clipping

Clipping bias bounded by Wasserstein distances

Gradient Distribution of NN

Visualization with random projection

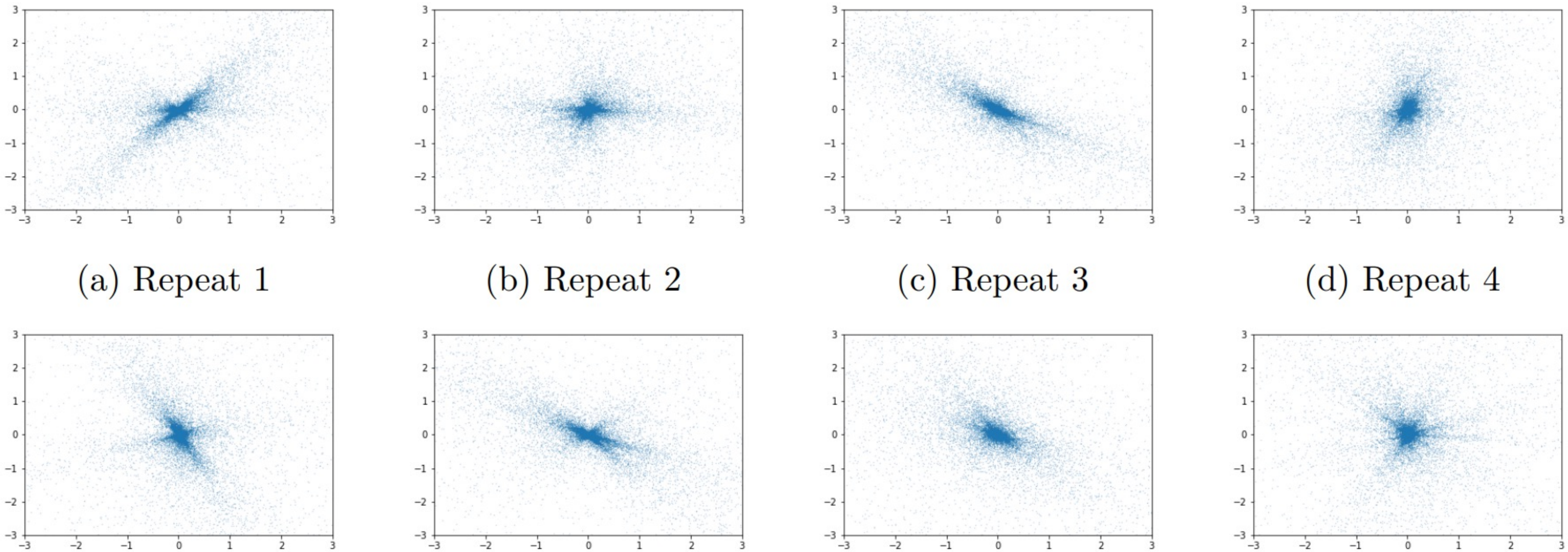


(a) Epoch 0 (b) Epoch 5 (c) Epoch 10 (d) Epoch 59

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).

Gradient Distribution of NN

Multiple random projections



(a) Repeat 1 (b) Repeat 2 (c) Repeat 3 (d) Repeat 4

(e) Repeat 5 (f) Repeat 6 (g) Repeat 7 (h) Repeat 8

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

Beyond Symmetric Distributions

- Positively skewed distributions:
  $\tilde{p}(\xi) \geq \tilde{p}(-\xi)$, for all ξ with $\langle \xi, \nabla f(x) \rangle$

- Certain mixtures of symmetric distributions

Clipping bias correction method:

- Adding pre-clipping noise

Part 2: Low-dimensional structure in gradients

- *Yingxue Zhou, Z. S.W., Arindam Banerjee*
  "Bypassing the Ambient Dimension: Private SGD with Gradient Subspace"
  In ICLR 2021

Dimensionality

Gradient norm bound

$$O\left(\frac{C\sqrt{d\ln(1/\delta)}}{n\epsilon}\right) + \frac{1}{T}\sum_{t=1}^{T} W_{\nabla f(x_t),C}(\tilde{p}_t, p_t)$$

DP-SGD without clipping
Depends on ambient dimension d

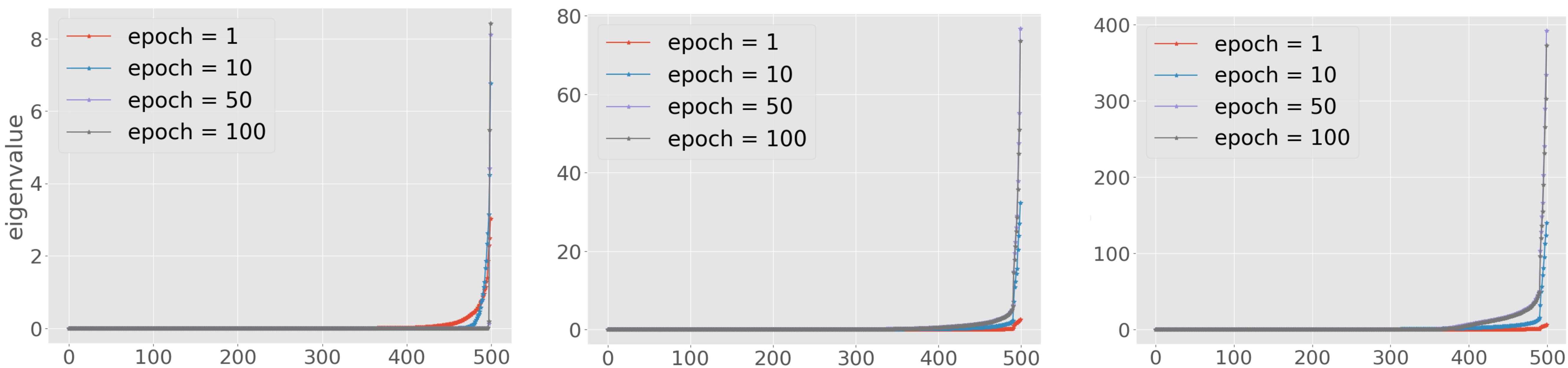
Clipping bias

Spectrum of Gradient Second Moments

Eigenvalues of

$$M_t = \mathbb{E}[\nabla \ell(x_t, s_i) \nabla \ell(x_t, s_i)^\intercal]$$

SGD   DP-SGD $\sigma = 1$ DP-SGD $\sigma = 2$



Order or eigenvalues from largest to smallest

Ambient dimension $d \approx 130{,}000$

[ZWB21]

Projected DP-SGD (PDP-SGD)

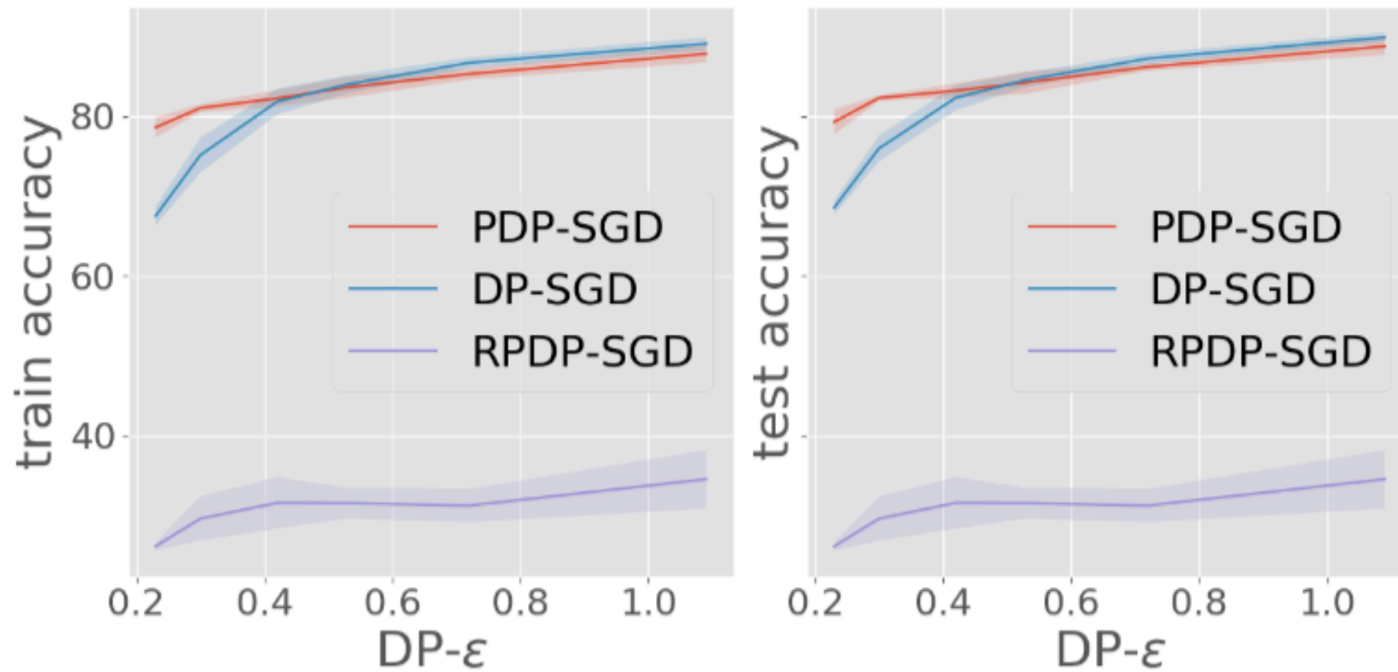
Assume small amount of public data (no privacy concern)

PDP-SGD [ZWB21]

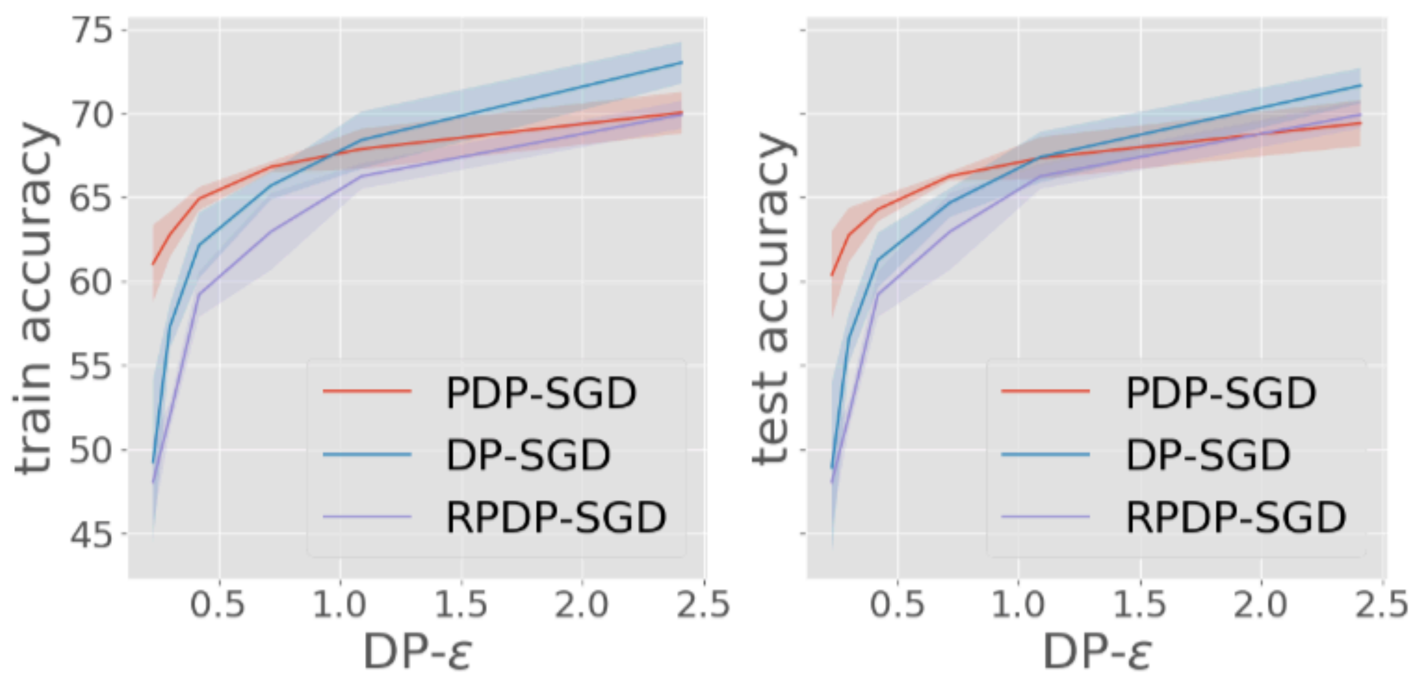
- For $t = 1,\ldots,T$
  - Gradient estimate on a mini-batch S_t :
    $$\tilde{g}_t = \left(\frac{1}{|S_t|} \sum_{i \in S_t} \nabla \ell(x_t; s_i) \right) + \mathcal{N}(0, \sigma^2 I)$$
  - Use public data to compute projection Π_k onto the top-k eigenspace of M_t
  - Update :
    $$x_{t+1} = x_t - \eta \, \Pi_k \tilde{g}_t$$

Balancing two sources of error

- Error due to projection
  $\|\Pi_k \nabla \ell(x; s_i) - \nabla \ell(x; s_i)\|$
- Gradient perturbation in the subspace $\approx \frac{\sqrt{k}}{n\epsilon}$
  (from $\sqrt{d}$ to $\sqrt{k}$)

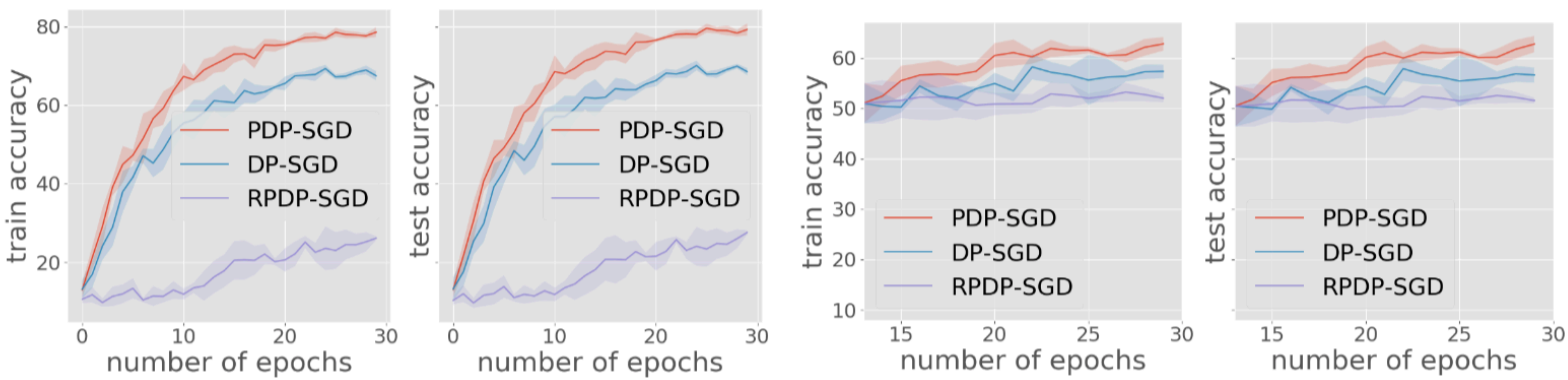


(a) MNIST



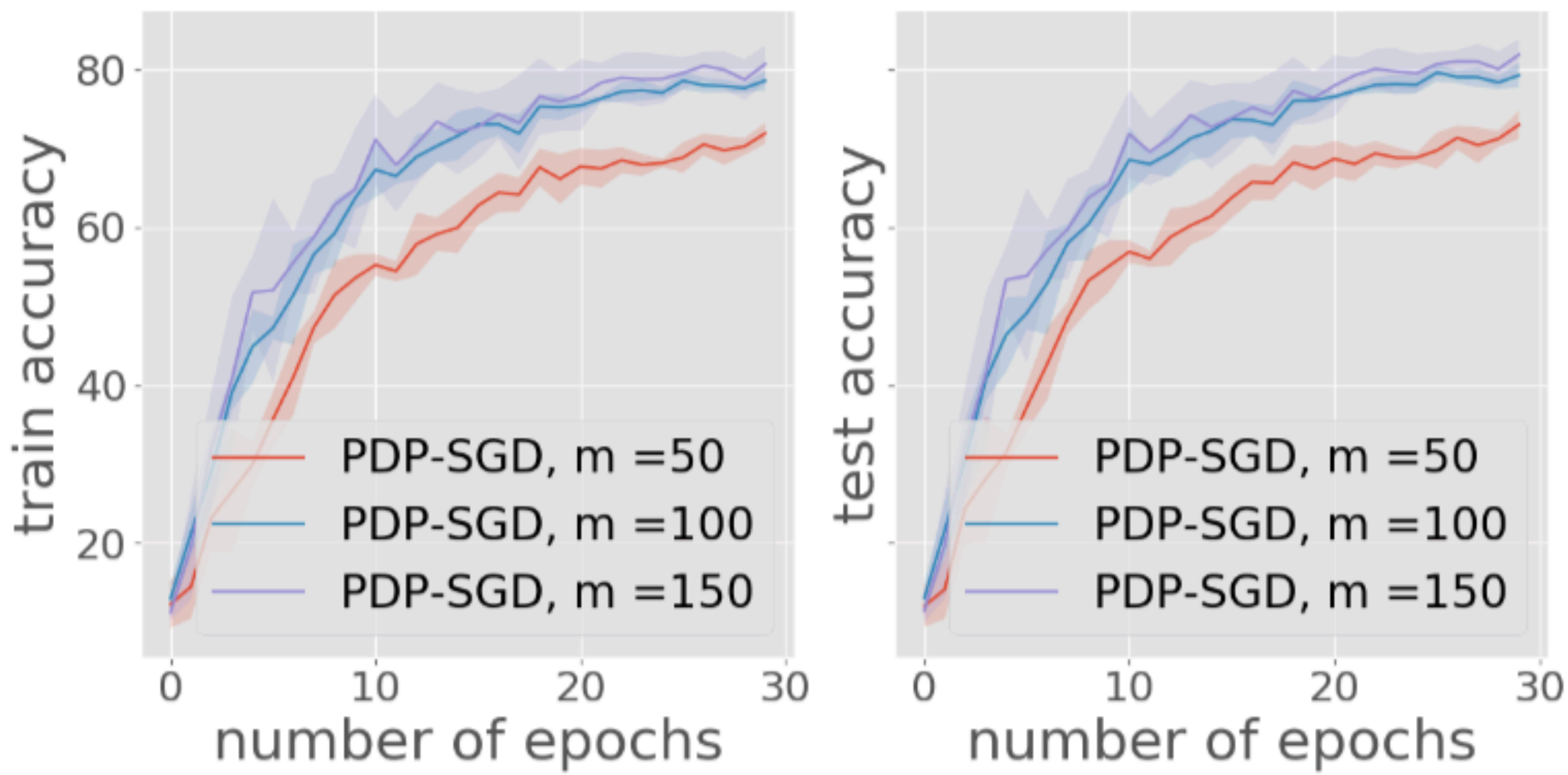
(b) Fashion MNIST

Training Dynamics

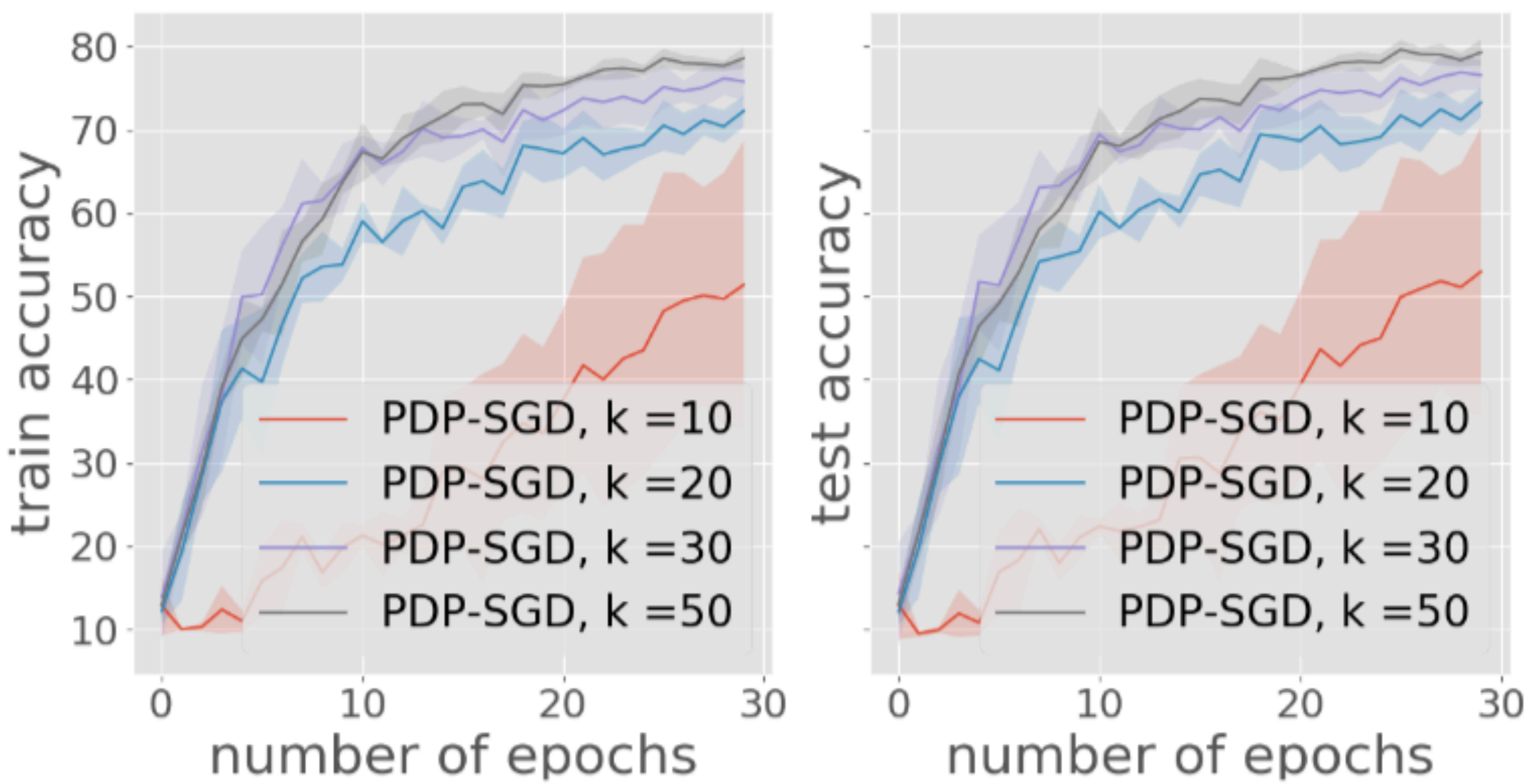


(a) MNIST, $\epsilon = 0.23$ (b) Fashion MNIST, $\epsilon = 0.30$

Size of public dataset m

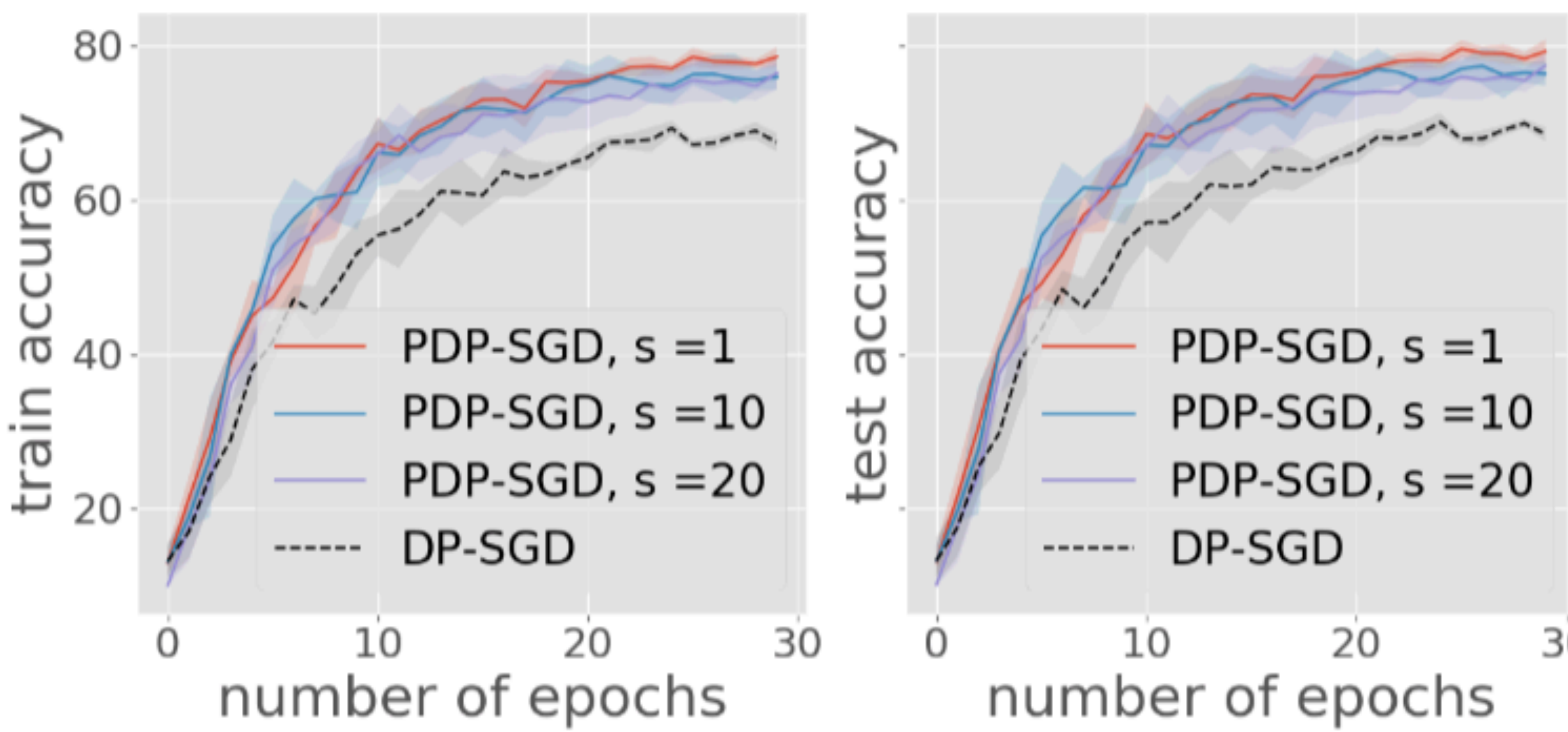


Different choices of k

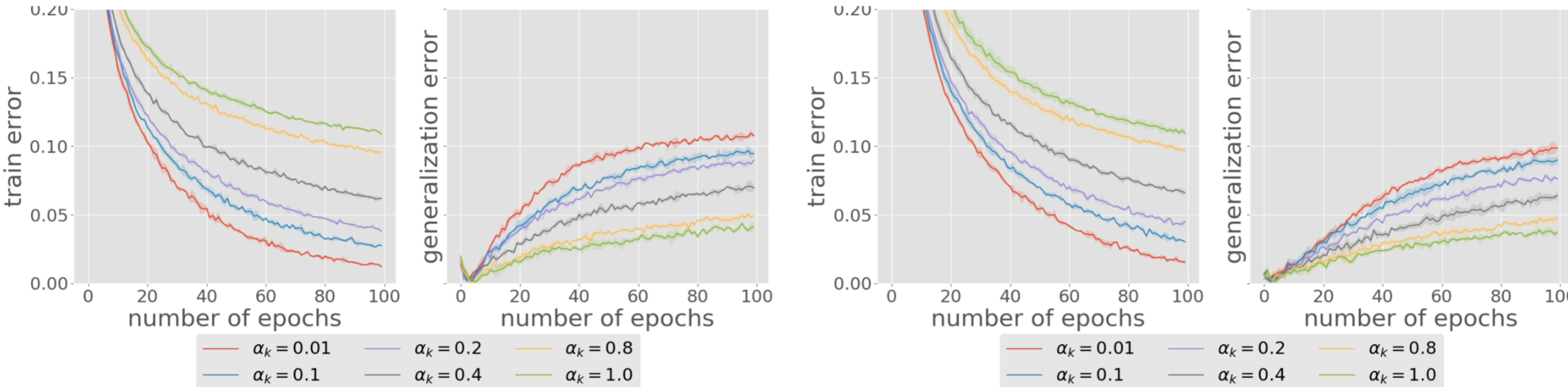


Frequency of computing subspace

Updating every s rounds



Stochastic Gradient Langevin Dynamics



(a) VGG-19

(b) ResNet-18

- α_k denotes noise rate