Loss of Noisy Stochastic Gradient Descent Might Converge Even for Non-Convex Losses

Shahab Asoodeh

Mario Diaz

McMaster University

UNAM

Noisy-SGD with hidden states

Dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$, learning parameter η, parameter space $\mathcal{W}$

- $W_0 \leftarrow$ random point in $\mathcal{W}$
- for $t = 1$ to T do
  - $B_t \leftarrow$ random mini-batch of size b
  - $W_t \leftarrow \Pi_{\mathcal{W}}\left(\psi_{B_t}(W_{t-1}) + \sigma^2 Z_t\right)$
- return W_T

only last update is released

$$\psi_B(w) \triangleq w - \frac{\eta}{b} \sum_{i \in B} \nabla \ell(w, x_i)$$

Feldman, Mironov, Talwar, and Thakurta, "Privacy Amplification by Iteration", 2018.

Previous work

Differential Privacy Dynamics of Langevin Diffusion and Noisy Gradient Descent

Rishav Chourasia[*], Jiayuan Ye[*], Reza Shokri
Department of Computer Science, National University of Singapore
{rishav1, jiayuan, reza}@comp.nus.edu.sg

Assumptions:

- Full batch: $b = n$
- $w \mapsto \ell(w, x)$ is λ-strongly convex for all x
- $w \mapsto \ell(w, x)$ is smooth for all x

Rényi differential privacy parameter of Noisy-GD converges as $T \to \infty$.

Previous work

Differentially Private Learning Needs Hidden State (Or Much Faster Convergence)

Jiayuan Ye, Reza Shokri
Department of Computer Science
National University of Singapore
{jiayuan, reza}@comp.nus.edu.sg

Assumptions:

- $w \mapsto \ell(w, x)$ is λ-strongly convex for all x
- $w \mapsto \ell(w, x)$ is smooth for all x

Privacy of Noisy Stochastic Gradient Descent: More Iterations without More Privacy Loss

Jason M. Altschuler
MIT
jasonalt@mit.edu

Kunal Talwar
Apple
ktalwar@apple.com

Assumptions:

- $w \mapsto \ell(w, x)$ is convex for all x
- $w \mapsto \ell(w, x)$ is smooth for all x

Rényi differential privacy parameter of Noisy-SGD converges as $T \to \infty$.

Main result

Question. Does differential privacy parameter of Noisy-SGD converge even for non-convex loss functions?

Answer. Yes, provided that gradients are clipped.

$$\psi_B(w) \triangleq w - \frac{\eta}{b} \sum_i \mathrm{Clip}(\nabla \ell(w, x_i))$$

$$\mathrm{Clip}(v) \triangleq \min\left\{1, \frac{C}{\|v\|}\right\} v$$

DP-SGD

- $W_0 \leftarrow$ random point in $\mathcal{W}$
- for $t = 1$ to T do
  - $B_t \leftarrow$ random mini-batch of size b
  - $W_t \leftarrow \Pi_{\mathcal{W}}\left(\psi_{B_t}(W_{t-1}) + \sigma^2 Z_t\right)$
- return W_T

Abadi, Chu, Goodfellow, McMahan, Mironov, Talwar, and Zhang, “Deep Learning with Differential Privacy”, 2016.

Main result

for an **arbitrary** loss function

[Informal]. The DP-SGD algorithm is (ε, δ)-DP with

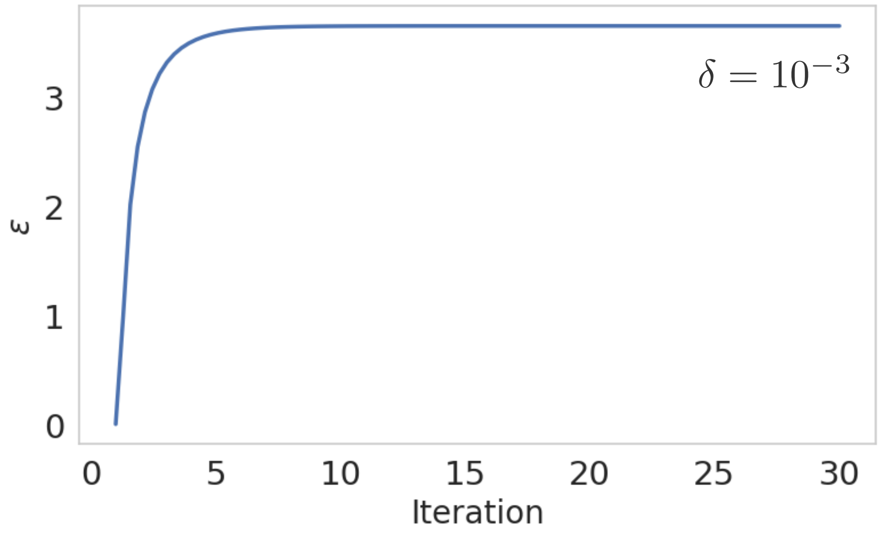
$$\delta \leq \frac{1 - [(1-p)\theta]^T}{1 - (1-p)\theta} \cdot p \cdot \theta,$$

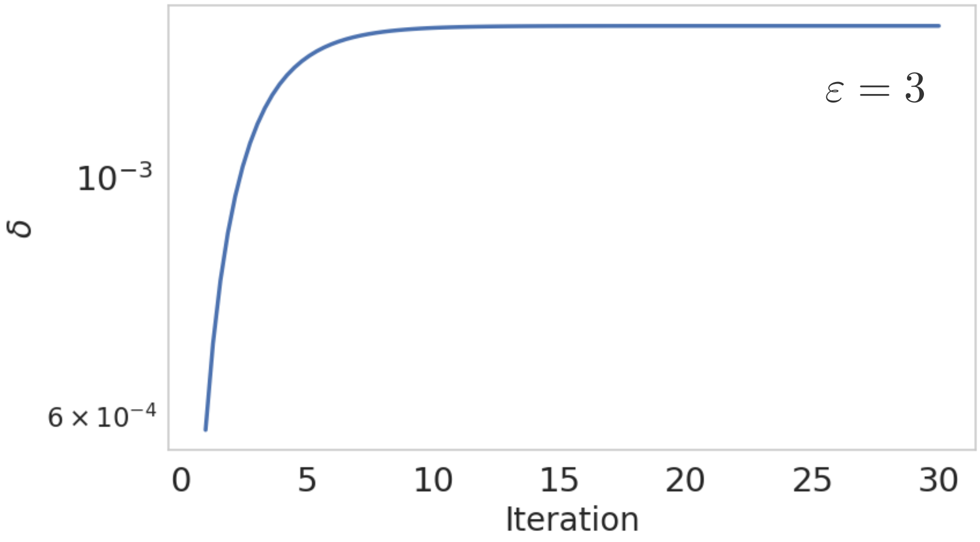
where $p = b/n$, and $\theta \in (0,1)$ is a constant depending on $\varepsilon, \eta, C, \sigma^2$, and $\text{dia}(\mathcal{W})$.

More formally, $\theta \triangleq \theta_\varepsilon\left(\frac{\text{dia}(\mathcal{W}) + 2\eta C}{\sigma}\right)$, where

$$\theta_\varepsilon(r) \triangleq Q\left(\frac{\varepsilon}{r} - \frac{r}{2}\right) - e^\varepsilon Q\left(\frac{\varepsilon}{r} + \frac{r}{2}\right).$$

<ins>Proof idea</ins>: coupled non-linear data processing inequality for Gaussian kernels





$\text{dia}(\mathcal{W}) = 1, C = 1, \eta = 0.01, \sigma = 1, p = 0.001$