Private SGD with Shuffling

Eloïse Berthier · Sai Praneeth Karimireddy

We study the Rényi differential privacy of **cyclic SGD**, with training examples sampled without replacement. We propose two examples where gradient-perturbed cyclic SGD is both **not faster** yet **less private** than plain SGD.

Differential Privacy for Machine Learning

Motivation

- Machine learning models can leak private information from their training data.
- This enables reconstruction or membership attacks that could be harmful in some applications: medicine, census, NLP…
- Ideally, the model should not depend too much on one individual data point.

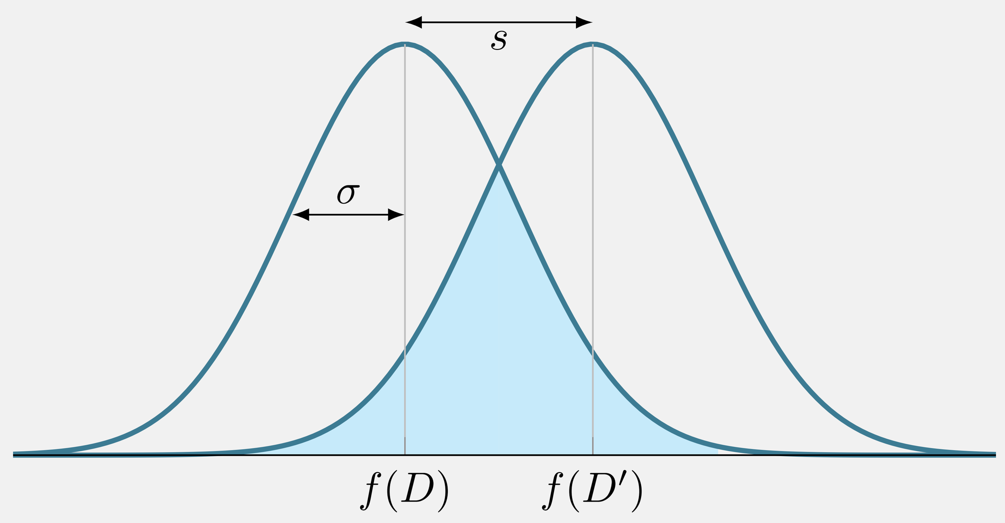
Consider two datasets D and D' differing only by one data point. A randomized algorithm outputs respectively p or q when run on each dataset. It has $(\alpha, \varepsilon(\alpha))$ RDP if the two probability distributions are statistically undistinguishable:

Rényi Differential Privacy

$$D_\alpha(p||q) := \frac{1}{\alpha-1}\log\int p(u)^\alpha q(u)^{1-\alpha}\mathrm{d}u \leq \varepsilon(\alpha)$$

Adding noise to the query: $f(D) + \mathcal{N}(0, \sigma^2 I_d)$

Ensures privacy: $\varepsilon(\alpha) = \dfrac{\alpha s^2}{2\sigma^2}$



Main Properties

1. Post-processing cannot increase ε.
2. Composition increases ε linearly with the number of queries: $\varepsilon \rightarrowtail t\varepsilon$
3. Subsampling shrinks ε quadratically by the subsampling parameter: $\varepsilon \rightarrowtail \rho^2\varepsilon$

Privacy of SGD with replacement

Gradient perturbation with noise $\eta_t \sim \mathcal{N}(0, \sigma^2 I_d)$

$$\theta_{t+1} = \theta_t - \gamma(\nabla f(\theta_t, x_{i(t)}) + \eta_t)$$

Assuming the gradients are bounded by L and $\sigma \geq 2L$, after one epoch, the algorithm achieves RDP:

$$\varepsilon(\alpha) \leq \mathcal{O}\left(n \times \frac{1}{n^2} \times \frac{\alpha L^2}{\sigma^2}\right)$$

composition subsampling

"Private" Cyclic SGD is Not Faster

Why Should I Shuffle?

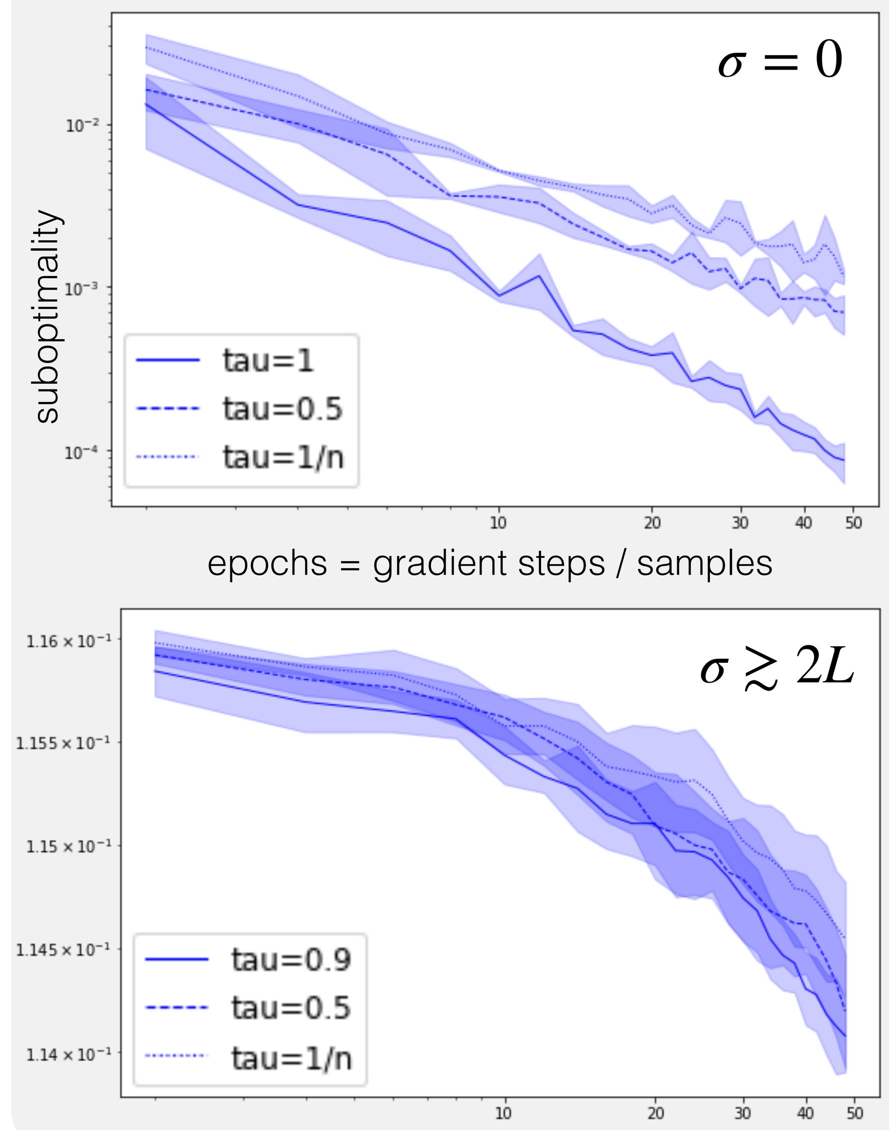
- It is what is used in practice.
- Allows faster implementations, including in distributed settings.
- It has been recently proved that it enjoys a faster convergence rate of $\mathcal{O}(1/T^2)$ against $\mathcal{O}(1/T)$ for plain SGD (on *strongly convex* objectives).

✦ However, no privacy guarantees are known for this algorithm because sampling is no longer independent.

Algorithm 1 Private SGD with Shuffling

procedure PSGD($\mathcal{D}, \theta_0, \gamma, \ell, L, e, \tau, \sigma^2$)
  $\theta \leftarrow \theta_0$
  for $i = 1$ to e/τ **do**
    $(s_1, \ldots, s_n) \leftarrow \mathit{Shuffle}(\{1, \ldots, n\})$
    for $j = 1$ to $\lfloor \tau n \rfloor$ **do**
      $x \leftarrow x_{s_j}$
      $g \leftarrow \mathit{Clip}(\nabla_\theta \ell(\theta, x), L)$
      Sample $\eta \sim \mathcal{N}(0, \sigma^2 I_d)$
      $\theta \leftarrow \theta - \gamma(g + \eta)$
  return θ

At the beginning of each epoch, shuffle the dataset. Then do one pass through, and stop after a fixed fraction has been seen. Start again.

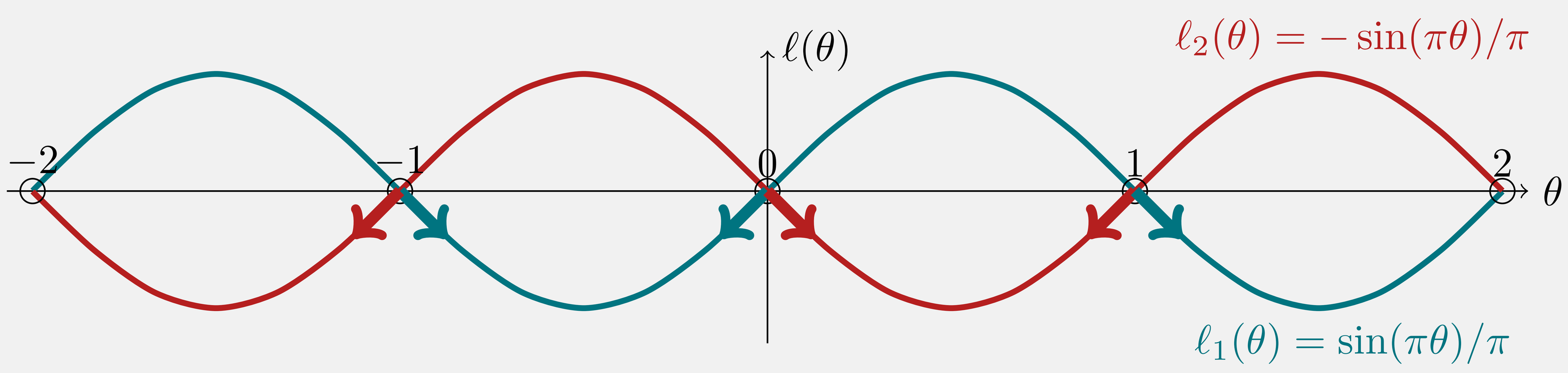


✔ Cyclic SGD is faster in a non-private, strongly convex setting.

✔ But not faster in a private *gradient perturbation* setting.

Shuffling can be Less Private

✦ Are privacy guarantees weaker for full cyclic SGD than for plain SGD ?



- Two possible loss functions $\ell_1(\theta)$ or $\ell_2(\theta)$
- Do two SGD steps with $\theta_0 = 1$ and $\gamma = 1$
- With probability 0.5, output final θ
- Else output one at random in $\{-2, 0, 2\}$

✔ Sampling with replacement is more private than without: $(\infty, \log(1.75))$ vs $(\infty, \log(4))$.

✔ Choosing twice the same data point preserves more privacy.

✔ The final iterate reveals the whole trajectory.

References

- Carlini, N., Liu, C., Kos, J., Erlingsson, Ú., & Song, D. (2018). The secret sharer: Measuring unintended neural network memorization & extracting secrets. *arXiv preprint arXiv:1802.08232.*
- Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science, 9*(3–4).
- Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conf. on Computer & Communications Security*
- Wang, Y. X., Balle, B., & Kasiviswanathan, S. (2018). Subsampled Rényi Differential Privacy and Analytical Moments Accountant. *arXiv preprint arXiv:1808.00087.*
- Haochen, J., & Sra, S. (2019, May). Random Shuffling Beats SGD after Finite Epochs. In *International Conference on Machine Learning* (pp. 2624-2633).