Dynamic Privacy Budget Allocation Improves Data Efficiency of Differentially Private Gradient Descent

Junyuan Hong\textsuperscript{1}, Zhangyang Wang\textsuperscript{2}, Jiayu Zhou\textsuperscript{1}

Michigan State University, University of Texas at Austin
Privacy Regulations and Risks

- **GDPR**: General Data Protection Regulation
- **HIPAA**: Health Insurance Portability and Accountability Act, 1996
- **SOX**: Sarbanes-Oxley Act, 2002
- **PCI**: Payment Card Industry Data Security Standard, 2004
- **SHIELD**: Stop Hacks and Improve Electronic Data
- **Security Act**, Jan 1 2019
Differential Privacy

\[ Z(y) \triangleq \log \left( \frac{p(\mathcal{A}(D) = y)}{p(\mathcal{A}(D') = y)} \right) \]

where \( y \sim \mathcal{A}(D) \) and \( D, D' \) are adjacent (differing at one sample)
Differentially Private Stochastic Gradient Descent (DPSGD)

- Non-private SGD: $\theta_{t+1} = \theta_t - \eta \nabla_t$
- Private SGD: $\theta_{t+1} = \theta_t - \eta g_t$, $g_t = \text{Privatize}(\nabla_t)$

**Algorithm 1** Privatizing gradients

**Input:** Private gradient $\nabla_t$ summed from $[\nabla_t^{(1)}, \ldots, \nabla_t^{(n)}]$, residual privacy budget $R_t$

1: $\bar{\nabla}_t \leftarrow \frac{1}{N} \sum_{n=1}^{N} \nabla_t^{(n)} \min \{1, C_t/\|\nabla_t^{(n)}\| \}$ \hspace{1cm} \triangleright \text{Sensitivity constraint}
2: $\rho_t \leftarrow 1/\sigma_t^2$
3: if $\rho_t < R_t$ then
4: \hspace{0.5cm} $R_{t+1} \leftarrow R_t - \rho_t$
5: \hspace{0.5cm} $g_t \leftarrow \bar{\nabla}_t + C_t\sigma_t\nu_t/N$, $\nu_t \sim N(0, I)$ \hspace{1cm} \triangleright \text{Privacy noise}
6: \hspace{0.5cm} return $\eta_t g_t, R_{t+1}$ \hspace{1cm} \triangleright \text{Utility projection}
7: else
8: \hspace{0.5cm} Terminate
DPSGD needs more data

| Algorithm                      | Schedule \((\sigma^2_t)\) | Utility Upper Bound                                                                 |
|-------------------------------|----------------------------|-------------------------------------------------------------------------------------|
| *GD+Adv [3]                   | \(O\left(\frac{\ln(N/\delta)}{R_{\epsilon,\delta}}\right)\) | \(O\left(\frac{D \ln^3 N}{N R_{\epsilon,\delta}}\right)\)                      |
| GD+MA [34]                    | \(O\left(\frac{T}{R_{\epsilon,\delta}}\right)\) | \(O\left(\frac{D \ln^2 N}{N^2 R_{\epsilon,\delta}}\right)\)                   |
| GD+MA (adjusted utility) [39] | \(O\left(\frac{T}{R_{\epsilon,\delta}}\right)\) | \(O\left(\min\left\{\frac{\sqrt{D}}{N R_{\epsilon,\delta}}, \frac{D \ln N}{N^2 R_{\epsilon,\delta}^2}\right\}\right)\) |
| *GD+Adv+BBImp [7]             | \(O\left(\frac{n^2 \ln(n/\delta)}{R_{\epsilon,\delta}}\right)\) | \(O_p\left(\frac{D^2 \ln^2 \left(1/p\right)}{R_{\epsilon,\delta} N^{1-c}}\right)\) |
| Adam+MA [42]                  | \(O\left(\frac{T}{R_{\epsilon,\delta}}\right)\) | \(O_p\left(\frac{\sqrt{D} \ln \left(N \delta e/(1-p)\right)}{N R_{\epsilon,\delta}}\right)\) |
| GD, Non-Private               | 0                          | \(O\left(\frac{D}{N^2 R}\right)\)                                                |

\[
\frac{\ln^3 N}{N}
\]

\[
\frac{1}{N}
\]

How?
A close look at the private convergence

- Not converge to the optimal
  - Finite iteration
  - Noise
- Improve the final iterate loss given a privacy budget:
  \[ \text{EER} = \mathbb{E}_\nu[f(\theta_{T+1})] - f(\theta^*) \]
  - The upper bound of EER

Strictly private

High variance and away from optima

Less private
Why study convergence upper bound?

- Bound the worst case (highest errors).
- Find a way to speed up optimization algorithm.
- Gain insights into privacy operations, e.g., noise magnitude, clipping norm, etc.
- To compare different algorithms: convergence rate.
Assumptions

- $G$-Lipschitz continuous loss,

\[ \| f(x) - f(x') \| \leq G \| x - x' \| \Leftrightarrow \| f'(x) \| \leq G \] if $f$ is differentiable.

- $M$-Lipschitz continuous gradient or $M$-smooth loss:

\[ \| \nabla f(x) - \nabla f(x') \| \leq M \| x - x' \| \]

- $\mu$-Polyak-Lojasiewicz (PL) condition < $\mu$-strongly convex

\[ \| \nabla f(\theta) \|^2 \geq 2\mu(f(\theta) - f(\theta^*)) \]
Revisit: Convergence of DPSGD with non-static $\sigma_t$

\textbf{Theorem 3.2.} Let $\alpha$, $\kappa$ and $\gamma$ be defined in Eq. (5), and $\eta_t = \frac{1}{M}$. Suppose $f(\theta; x_t)$ is $G$-Lipschitz $M$-smooth and satisfies the Polyak-Lojasiewicz condition. If $C_t \leq G$, then clipping does not take place, i.e., $\tilde{\nabla}_t = \nabla_t$ and the following holds:

$$EER = E_{\nu}[f(\theta_{T+1})] - f(\theta^*) \leq \left( \gamma^T + R \sum_{t=1}^{T} q_t \sigma_t^2 \right) (f(\theta_1) - f(\theta^*)),$$

where $q_t \triangleq \gamma^{T-t} \alpha_t$. \hspace{1cm} (6)

$$\alpha_t \triangleq \frac{MD}{2R} \left( \frac{\eta_t C_t}{N} \right)^2 \frac{1}{f(\theta_1) - f(\theta^*)} > 0, \ \kappa \triangleq \frac{M}{\mu} \geq 1, \text{ and } \gamma \triangleq 1 - \frac{1}{\kappa} \in [0, 1). \hspace{1cm} (5)$$
Revisit: Convergence of DPSGD with non-static $\sigma_t$

Theorem 3.2. Let $\alpha$, $\kappa$ and $\gamma$ be defined in Eq. (5), and $\eta_t = \frac{1}{M}$. Suppose $f(\theta; x_t)$ is $G$-Lipschitz $M$-smooth and satisfies the Polyak-Lojasiewicz condition. If $C_t \leq G$, then clipping does not take place, i.e., $\tilde{\nabla}_t = \nabla_t$ and the following holds:

$$EER \leq \left( \gamma^T + R \sum_{t=1}^{T} q_t \sigma_t^2 \right) (f(\theta_t) - f(\theta^*)), \quad (6)$$

where $q_t = \gamma^{T-t} \alpha_t$. \hfill (7)

- Finite iteration
- Noise impact

- Schedule noise to
  - Extend iteration $T$
  - Reduce the effect of noise
Revisit: Convergence of DPSGD with non-static $\sigma_t$

**Theorem 3.2.** Let $\alpha$, $\kappa$, and $\gamma$ be defined in Eq. (5), and $\eta_t = \frac{1}{M}$. Suppose $f(\theta; x_t)$ is $G$-Lipschitz $M$-smooth and satisfies the Polyak-Lojasiewicz condition. If $C_t \leq G$, then clipping does not take place, i.e., $\tilde{\nabla}_t = \nabla_t$ and the following holds:

$$
\text{EER} \leq \left( \gamma^T + R \sum_{t=1}^{T} q_t \sigma_t^2 \right) (f(\theta_t) - f(\theta^*)),
$$

where $q_t \triangleq \gamma^{T-t} \alpha$. (6)

Influence of noise

**Lemma 3.1 (Dynamic schedule).** Suppose $\sigma_t$ satisfy $\sum_{t=1}^{T} \sigma^{-2} = R$. Given a positive sequence $\{q_t\}$, the following equation holds

$$
\min_{\sigma} R \sum_{t=1}^{T} q_t \sigma_t^2 = \left( \sum_{t=1}^{T} \sqrt{q_t} \right)^2, \quad \text{when } \sigma_t = \sqrt{\frac{1}{R \sum_{i=1}^{T} \sqrt{\frac{q_i}{q_t}}}}.
$$

Reduce noise impact

How much improvement can we achieve?
Advantage of dynamic schedule

Theorem 3.3. When $\sigma_t = \sqrt{T/R}$ and $C_t$ be constant, let $\alpha = \alpha_t$, $\gamma$ and $\kappa$ be defined in Eq. (5) and the $T$ minimizing the upper bound of Eq. (6) is\(^1\)

$$T^{\text{uniform}} = \begin{cases} \left\lfloor \log_\gamma \left( \frac{\kappa \alpha}{\ln(1/\gamma)} \right) \right\rfloor & \kappa \alpha + \ln \gamma < 0 \\ \kappa \alpha + \ln \gamma \geq 0 & \end{cases} \quad (8)$$

Meanwhile, for $\kappa > 1$, the minimal bound is

$$\text{ERUB}_{\text{min}}^{\text{uniform}} = \begin{cases} \Theta \left( \kappa^2 \alpha \left[ 1 + (\kappa^2 \alpha - 1) \ln(\kappa^2 \alpha) \right] \right) & \kappa \alpha + \ln \gamma < 0 \\ 1 & \kappa \alpha + \ln \gamma \geq 0 \end{cases} \quad (9)$$

non-private ERUB : $\alpha \triangleq \frac{DG^2}{2RMN^2(f(\theta_1) - f(\theta^*))} \leq O \left( \frac{DG^2}{RMN^2} \right)$, \quad (4)

curvature : $\kappa \triangleq \frac{M}{\mu}$, \quad (5)

convergence rate : $\gamma \triangleq 1 - \frac{1}{\kappa}$, \quad (6)
Advantage of dynamic schedule

Theorem 3.3. When $\sigma_t = \sqrt{T/R}$ and $C_t$ be constant, let $\alpha = \alpha_t$, $\gamma$ and $\kappa$ be defined in Eq. (5) and the $T$ minimizing the upper bound of Eq. (6) is

$$T^\text{uniform} = \left\{ \begin{array}{ll}
\log_\gamma \left( \frac{\kappa \alpha}{\ln(1/\gamma)} \right), & \kappa \alpha + \ln \gamma < 0 \\
0, & \kappa \alpha + \ln \gamma \geq 0
\end{array} \right. \quad (8)$$

Meanwhile, for $\kappa > 1$, the minimal bound is

$$\text{ERUB}_{\text{min}}^\text{uniform} = \left\{ \begin{array}{ll}
\Theta \left( \kappa^2 \alpha \left[ 1 + (\kappa^2 \alpha - 1) \ln(\kappa^2 \alpha) \right] \right), & \kappa \alpha + \ln \gamma < 0 \\
1, & \kappa \alpha + \ln \gamma \geq 0
\end{array} \right. \quad (9)$$

Lemma 3.2. Let $\alpha$, $\kappa$ and $\gamma$ be defined in Eq. (5). When $\sigma_t$ be defined as Eq. (10), the $T$ minimizing the upper bound of Eq. (6) is

$$T^* = \left[ 2 \log_\gamma \left( \frac{\alpha}{\alpha + (1 - \sqrt{\gamma})^2} \right) \right]. \quad (11)$$

Meanwhile, the minimal bound is

$$\text{ERUB}_{\text{min}}^\text{dynamic} = \Theta \left( \frac{\kappa^2 \alpha}{\kappa^2 \alpha + 1} \right). \quad (12)$$
Advantage of dynamic schedule on optimal upper bound

# of allowed iterations

![Graph showing the number of allowed iterations for different loss curvatures.](image)

**Smooth loss curvature**

**Sharp loss curvature**

**Excess Expected Risks**

![Graph showing excess expected risks for different loss curvatures.](image)

**stable when the loss curvature (κ) is sharp**
Advantage of dynamic schedule

- Empirically check the $q_t$

$$
EER \leq \left( \gamma^T + R \sum_{t=1}^{T} q_t \sigma_t^2 \right) (f(\theta_1) - f(\theta^*)) ,
$$
where $q_t \triangleq \gamma^{T-t} \alpha_t$.
Further reduce the noise by momentum

- Example of momentum in modern optimizers: Adam, SGD with momentum

**Algorithm 2** Privatizing gradients with debiased momentum

**Input:** Private gradient $\nabla_t$ summed from $[\nabla_t^{(1)}, \ldots, \nabla_t^{(N)}]$, residual privacy budget $R_t$

1. $\tilde{\nabla}_t \leftarrow \frac{1}{N} \sum_{n=1}^{N} \nabla_t^{(n)} \min\{1, C_t / \| \nabla_t^{(n)} \| \}$  \hspace{1cm} ▶ Sensitivity constraint
2. $\rho_t \leftarrow 1 / \sigma_t^2$
3. if $\rho_t < R_t$ then
4. \hspace{.5cm} $R_{t+1} \leftarrow R_t - \rho_t$
5. $g_t \leftarrow \tilde{\nabla}_t + \nu_t, \nu_t \sim \mathcal{N}(0, (C_t \sigma_t/N)^2 I)$  \hspace{1cm} ▶ Privacy noise
6. $v_{t+1} = \beta v_t + (1 - \beta) g_t, \; v_1 = 0$
7. $\hat{v}_{t+1} = v_{t+1} / (1 - \beta^t)$
8. return $\eta_t \hat{v}_{t+1}, R_{t+1}$  \hspace{1cm} ▶ Utility projection
9. else
10. Terminate
Further reduce the noise by momentum

**Theorem 3.4** (Convergence under PL condition). Suppose $f(\theta; x_t)$ is $M$-smooth, $G$-Lipschitz and satisfies the Polyak-Lojasiewicz condition. Let $\eta_t = \eta_0$. If $C_t \geq G$ which implies $\nabla \theta_t = \nabla_i$ (clipping does not take place), then the following holds:

$$
\text{EER} \leq \gamma^T (f(\theta_1) - f(\theta^*)) + \frac{2\eta_0 D}{N^2} \sum_{t=1}^{T} q_t (C_t \sigma_t)^2 + \eta_0 \zeta \sum_{t=1}^{T} \gamma^{T-t} \|v_{i+1}\|^2
$$

(16)

where $q_t = \frac{\beta^2 (T-t+1) - \gamma (T-t+1)}{\beta^2 - \gamma}$, $\gamma = 1 - \eta_0 \mu$, $\zeta = \frac{4M^2 \beta \gamma}{(\gamma - \beta)^2 (1 - \beta)^3} \eta_0^2 + \frac{1}{2} M \eta_0 - 1$.  

(17)

Especially, when $\eta_0 \leq \frac{\beta (1-\beta)^3}{8M} \left[ \sqrt{\frac{1}{4} + \frac{16}{\beta (1-\beta)^3}} - 1 \right]$, the noise variance dominates the bound, i.e.,

$$
\text{EER} = \mathcal{O} \left( \frac{2\eta_0 D}{N^2} \sum_{t=1}^{T} q_t (C_t \sigma_t)^2 \right).
$$

A negative term if $\eta_0$ is small.

The GD noise

Proof partially based on (Zhu, et al., ArXiv 2020)
# Conclusion

| Algorithm                      | Schedule ($\sigma^2$) | Utility Upper Bound |
|-------------------------------|-----------------------|---------------------|
| *GD+Adv [3]                   | $O\left(\frac{\ln(N/\delta)}{R_c,\delta}\right)$ | $O\left(\frac{D\ln^3 N}{NR_c,\delta}\right)$ |
| GD+MA [34]                    | $O\left(\frac{T}{R_c,\delta}\right)$ | $O\left(\frac{D\ln^2 N}{N^2 R_c,\delta}\right)$ |
| GD+MA (adjusted utility) [39] | $O\left(\frac{T}{R_c,\delta}\right)$ | $O\left(\min\left(\frac{\sqrt{D}}{NR_c,\delta}, \frac{D\ln N}{N^2 R_c,\delta}\right)\right)$ |
| *GD+Adv+BBImp [7]             | $O\left(\frac{n^2 \ln(n/\delta)}{R_c,\delta}\right)$ | $O_p\left(\frac{D^2 \ln^2 (1/p)}{R_c,\delta N^{1-c}}\right)$ |
| Adam+MA [42]                  | $O\left(\frac{T}{R_c,\delta}\right)$ | $O_p\left(\frac{\sqrt{D} \ln (ND_{cf}(1-p))}{NR_c,\delta}\right)$ |
| GD, Non-Private               | 0                     | $O\left(\frac{D}{N^2 R}\right)$ |
| GD+zCDP, Static Schedule      | $\frac{T}{R}$         | $O\left(\frac{D \ln N}{N^2 R}\right)$ |
| GD+zCDP, Dynamic Schedule     | $O\left(\frac{(1-T/2)}{R}\right)$ | $O\left(\frac{D}{N^2 R}\right)$ |
| Momentum+zCDP, Static Schedule| $\frac{T}{R}$         | $O\left(\frac{D}{N^2 R} (c + \ln N \beta_{T>T})\right)$ |
| Momentum+zCDP, Dynamic Schedule| $O\left(\frac{c_1 Y^{T+t} + c_2 Y^{T-t}}{R}\right)$ | $O\left(\frac{D}{N^2 R} (1 + \frac{c D}{N^2 R} T_{T>T})\right)$ |

Improved sample efficiency approaching upper bound
How to estimate privacy policies?

• Learning to protect (Hong, et al. 2021): Transfer the dynamic policies learned from auxiliary tasks to private tasks based on the two insights:
  • Adaptive noise magnitude (this work)
  • Adaptive gradient sensitivity (Pichapati et al. 2019)
Thank you for your time!

Acknowledgments
This material is based in part upon work supported by National Institute of Aging (1RF1AG072449), Office of Naval Research (N00014-20-1-2382), National Science Foundation (IIS-1749940). Z. Wang is in part supported by Good Systems, a UT Austin Grand Challenge to develop responsible AI technologies.