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

Junyuan Hong[1], Zhangyang Wang[2], Jiayu Zhou[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

HIREVUE ASSESSMENTS

HireVue leverages artificial intelligence, video, as well as game-based and coding challenges to collect key candidate insights, enabling organizations to make more informed hiring decisions. HireVue Assessments are customizable to your hiring objectives or ready to deploy based on pre-validated models.

Learn More About HireVue Pre-employment Assessments >

https://www.hirevue.com/products/assessments

An algorithm can predict human behavior better than humans

https://qz.com/527008/an-algorithm-can-predict-human-behavior-better-than-humans/

Toronto police have been using facial recognition technology for more than a year — a tool police say increases the speed and efficiency of criminal investigations and has led to arrests in major crimes including homicides.

https://www.thestar.com/news/gta/2019/05/28/toronto-police-chief-releases-report-on-use-of-facial-recognition-technology.html

ROBOSTOP **Facebook shuts off AI experiment after two robots begin speaking in their OWN language only they can understand**

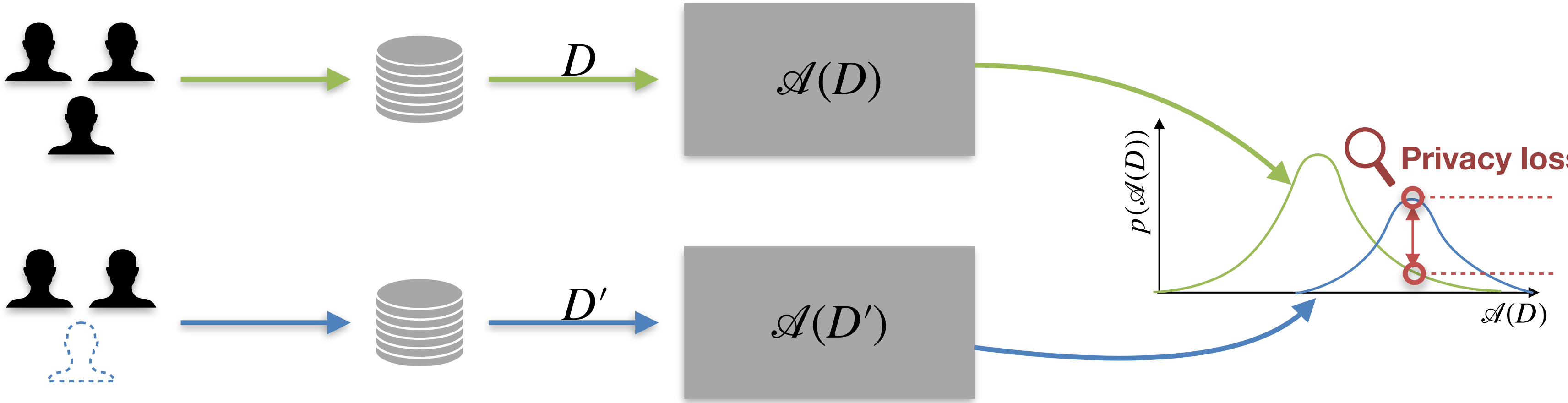
Experts have called the incident exciting but also incredibly scary

https://www.thesun.co.uk/tech/4141624/facebook-robots-speak-in-their-own-language/

IBM artificial intelligence can predict with 95% accuracy which workers are about to quit their jobs

https://www.cnbc.com/2019/04/03/ibm-ai-can-predict-with-95-percent-accuracy-which-employees-will-quit.html

Differential Privacy



Privacy loss at y $$Z(y) \triangleq \log\left(\frac{p(\mathscr{A}(D) = y)}{p(\mathscr{A}(D') = y)}\right)$$

where $y \sim \mathscr{A}(D)$ 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 ∇_t summed from $[\nabla_t^{(1)}, \dots, \nabla_t^{(n)}]$, residual privacy budget R_t

1: $\tilde{\nabla}_t \leftarrow \frac{1}{N} \sum_{n=1}^{N} \nabla_t^{(n)} \min\{1, C_t / \left\| \nabla_t^{(n)} \right\|\}$ ▷ Sensitivity constraint
2: $\rho_t \leftarrow 1/\sigma_t^2$ ▷ Budget request
3: **if** $\rho_t < R_t$ **then**
4: $\quad R_{t+1} \leftarrow R_t - \rho_t$
5: $\quad g_t \leftarrow \tilde{\nabla}_t + C_t \sigma_t \nu_t / N,\ \nu_t \sim \mathcal{N}(0, I)$ ▷ Privacy noise
6: $\quad$ **return** $\eta_t g_t, R_{t+1}$ ▷ Utility projection
7: **else**
8: $\quad$ Terminate

DPSGD needs more data

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

$$\frac{\ln^3 N}{N}$$

How?

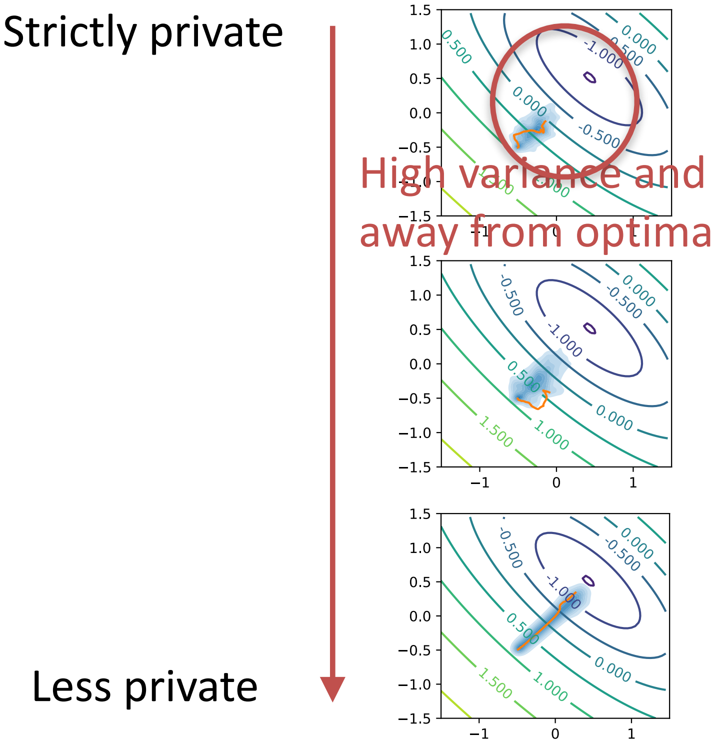
$$\frac{1}{N}$$

A close look at the private convergence

- Not converge to the optimal
  - Finite iteration
  - Noise
- Improve the final iterate loss given a privacy budget:

$$\mathrm{EER} = \mathbb{E}_{\nu}[f(\theta_{T+1})] - f(\theta^*)$$

  - The upper bound of EER



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,

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

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

  $\left\| \nabla f(x) - \nabla f(x') \right\| \leq M\|x - x'\|$

- μ-Polyak-Lojasiewicz (PL) condition < μ-strongly convex

  $\left\| \nabla f(\theta) \right\|^2 \geq 2\mu(f(\theta) - f(\theta^*))$

Revisit: Convergence of DPSGD with non-static σ_t

Theorem 3.2. *Let α, κ and γ be defined in Eq. (5), and $\eta_t = \frac{1}{M}$. Suppose $f(\theta; x_i)$ 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:*

$$\mathrm{EER} = \mathbb{E}_\nu[f(\theta_{T+1})] - f(\theta^*) \leq \left(\gamma^T + R\sum_{t=1}^{T} q_t \sigma_t^2\right)(f(\theta_1) - f(\theta^*)), \tag{6}$$

$$\text{where } q_t \triangleq \gamma^{T-t}\alpha_t. \tag{7}$$

$$\alpha_t \triangleq \frac{MD}{2R}\left(\frac{\eta_t C_t}{N}\right)^2 \frac{1}{f(\theta_1) - f(\theta^*)} > 0,\ \kappa \triangleq \frac{M}{\mu} \geq 1,\ \text{and}\ \ \gamma \triangleq 1 - \frac{1}{\kappa} \in [0, 1). \tag{5}$$

Proof partially based on (Wang, et al., NeurIPS 2017)

Revisit: Convergence of DPSGD with non-static σ_t

Theorem 3.2. *Let α, κ and γ be defined in Eq. (5), and $\eta_t = \frac{1}{M}$. Suppose $f(\theta; x_i)$ 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:*

$$\text{EER} \leq \left(\boxed{\gamma^T} + R \boxed{\sum_{t=1}^{T} q_t \sigma_t^2} \right) (f(\theta_1) - f(\theta^*)), \tag{6}$$

$$\text{where } q_t \triangleq \gamma^{T-t} \alpha_t. \tag{7}$$

Finite iteration

Noise impact

- Schedule noise to
  - Extend iteration T
  - Reduce the effect of noise

Revisit: Convergence of DPSGD with non-static σ_t

Theorem 3.2. *Let α, κ and γ be defined in Eq. (5), and $\eta_t = \frac{1}{M}$. Suppose $f(\theta; x_i)$ 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:*

$$\mathrm{EER} \leq \left(\gamma^T + R\boxed{\sum_{t=1}^T q_t \sigma_t^2}\right)(f(\theta_1) - f(\theta^*)), \tag{6}$$

$$\text{where } \boxed{q_t \triangleq \gamma^{T-t}\alpha_t}. \tag{7}$$

Influence of noise

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

✓Reduce noise impact

$$\min_\sigma R\sum_{t=1}^T q_t\sigma_t^2 = \left(\sum_{t=1}^T \sqrt{q_t}\right)^2, \text{ when } \sigma_t = \sqrt{\frac{1}{R}\sum_{i=1}^T \sqrt{\frac{q_i}{q_t}}}. \tag{10}$$

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 $\alpha = \alpha_t$, γ and κ be defined in Eq. (5) and the T minimizing the upper bound of Eq. (6) is*[1]

$$T^{*\,uniform} = \begin{cases} \left\lceil \log_\gamma \left(\frac{\kappa\alpha}{\ln(1/\gamma)} \right) \right\rceil, & \kappa\alpha + \ln\gamma < 0 \\ 0, & \kappa\alpha + \ln\gamma \geq 0 \end{cases} \tag{8}$$

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

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

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

$$\text{curvature} : \kappa \triangleq \frac{M}{\mu}, \tag{5}$$

$$\text{convergence rate} : \gamma \triangleq 1 - \frac{1}{\kappa}, \tag{6}$$

Advantage of dynamic schedule

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

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

$$\text{convergence rate}: \gamma \triangleq 1 - \frac{1}{\kappa}, \quad (6)$$

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

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

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

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

Lemma 3.2. *Let α, κ and γ be defined in Eq. (5). When σ_t be defined as Eq. (10), the T minimizing the upper bound of Eq. (6) is*

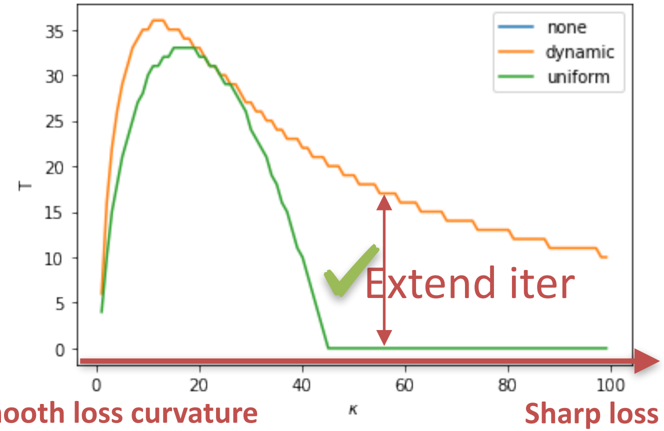
$$T^* = \left\lceil 2\log_\gamma \left(\frac{\alpha}{\alpha + (1 - \sqrt{\gamma})^2} \right) \right\rceil . \quad (11)$$

Meanwhile, the minimal bound is

$$\mathrm{ERUB}^{dynamic}_{\min} = \boxed{\Theta\left(\frac{\kappa^2\alpha}{\kappa^2\alpha + 1}\right).} \quad (12)$$

Advantage of dynamic schedule on optimal upper bound

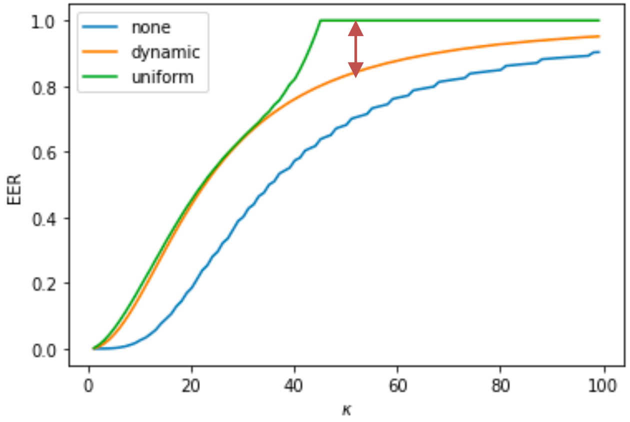
of allowed iterations



Smooth loss curvature

Sharp loss curvature

Excess Expected Risks



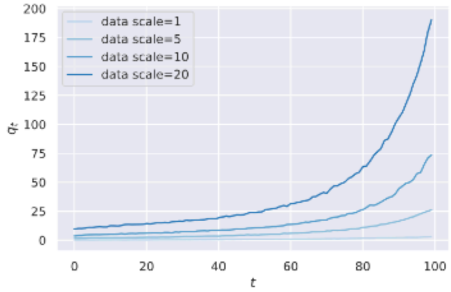
stable when the loss curvature (κ) is sharp

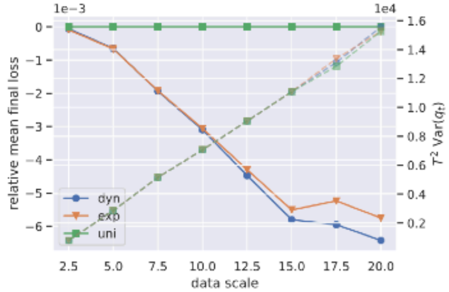
Advantage of dynamic schedule

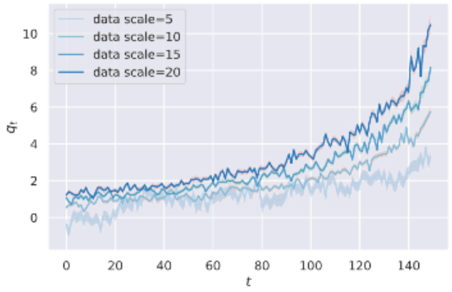
- Empirically check the q_t

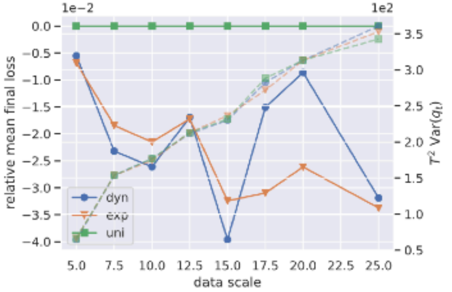
$$\mathrm{EER} \leq \left(\gamma^T + R\sum_{t=1}^{T} q_t\sigma_t^2\right)(f(\theta_1) - f(\theta^*)),$$

$$\textit{where } q_t \triangleq \gamma^{T-t}\alpha_t.$$









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 ∇_t summed from $[\nabla_t^{(1)}, \ldots, \nabla_t^{(n)}]$, residual privacy budget R_t

$\tilde{\nabla}_t \leftarrow \frac{1}{N}\sum_{n=1}^{N} \nabla_t^{(n)} \min\{1, C_t / \left\|\nabla_t^{(n)}\right\|\}$ ▷ Sensitivity constraint

$\rho_t \leftarrow 1/\sigma_t^2$ ▷ Budget request

if $\rho_t < R_t$ **then**

    $R_{t+1} \leftarrow R_t - \rho_t$

    $g_t \leftarrow \tilde{\nabla}_t + \nu_t,\ \nu_t \sim \mathcal{N}(0, (C_t\sigma_t/N)^2 I)$ ▷ Privacy noise

    $v_{t+1} = \beta v_t + (1-\beta) g_t,\ v_1 = 0$

    $\hat{v}_{t+1} = v_{t+1}/(1-\beta^t)$

    return $\eta_t \hat{v}_{t+1}, R_{t+1}$ ▷ Utility projection

else

    Terminate

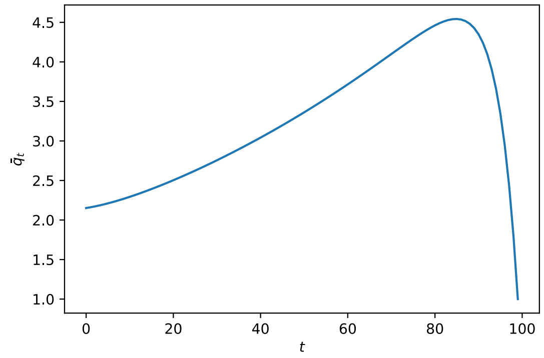
Further reduce the noise by momentum

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

$$\mathrm{EER} \leq \gamma^T (f(\theta_1) - f(\theta^*)) + \boxed{\frac{2\eta_0 D}{N^2} \underbrace{\sum_{t=1}^{T} q_t \left(C_t \sigma_t\right)^2}_{\textit{noise varinace}}} + \boxed{\eta_0 \zeta \underbrace{\sum_{t=1}^{T} \gamma^{T-t} \left\|v_{t+1}\right\|^2}_{\textit{momentum effect}}} \qquad (16)$$

$$\textit{where } q_t = \boxed{\frac{\beta^{2(T-t+1)} - \gamma^{T-t+1}}{\beta^2 - \gamma}}, \; \gamma = 1 - \eta_0 \mu, \; \zeta = \frac{4M^2 \beta \gamma}{(\gamma - \beta)^2 (1-\beta)^3} \eta_0^2 + \frac{1}{2} M \eta_0 - 1. \qquad (17)$$

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



A negative term if η_0 is small.

The GD noise

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

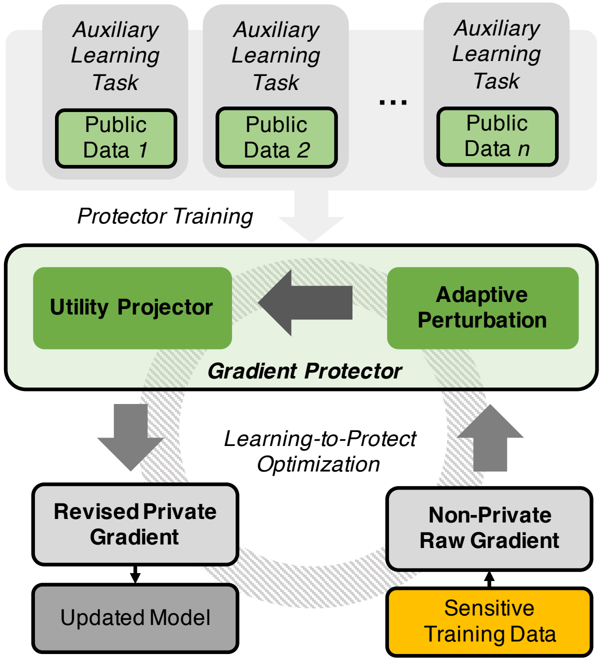
Conclusion

Algorithm	Schedule (σ_t^2)	Utility Upper Bound
*GD+Adv [3]	$\mathcal{O}\left(\frac{\ln(N/\delta)}{R_{\epsilon,\delta}}\right)$	$\mathcal{O}\left(\frac{D\ln^3 N}{NR_{\epsilon,\delta}}\right)$
GD+MA [34]	$\mathcal{O}(\frac{T}{R_{\epsilon,\delta}})$	$\mathcal{O}\left(\frac{D\ln^2 N}{N^2R_{\epsilon,\delta}}\right)$
GD+MA (adjusted utility) [39]	$\mathcal{O}(\frac{T}{R_{\epsilon,\delta}})$	$\mathcal{O}\left(\min \frac{\sqrt{D}}{NR_{\epsilon,\delta}}, \frac{D\ln N}{N^2R_{\epsilon,\delta}^2}\right)$
*GD+Adv+BBImp [7]	$\mathcal{O}\left(\frac{n^2\ln(n/\delta)}{R_{\epsilon,\delta}}\right)$	$\mathcal{O}_p\left(\frac{D^2\ln^2(1/p)}{R_{\epsilon,\delta}N^{1-c}}\right)$
Adam+MA [42]	$\mathcal{O}(\frac{T}{R_{\epsilon,\delta}})$	$\mathcal{O}_p\left(\frac{\sqrt{D}\ln(ND\epsilon/(1-p))}{NR_{\epsilon,\delta}}\right)$
GD, Non-Private	0	$\boxed{\mathcal{O}\left(\frac{D}{N^2R}\right)}$
GD+zCDP, Static Schedule	$\frac{T}{R}$	$\mathcal{O}\left(\frac{D\ln N}{N^2R}\right)$
GD+zCDP, Dynamic Schedule	$\mathcal{O}\left(\frac{\gamma^{(t-T)/2}}{R}\right)$	$\boxed{\mathcal{O}\left(\frac{D}{N^2R}\right)}$
Momentum+zCDP, Static Schedule	$\frac{T}{R}$	$\mathcal{O}\left(\frac{D}{N^2R}(c+\ln N\mathbb{I}_{T>\hat{T}})\right)$
Momentum+zCDP, Dynamic Schedule	$\mathcal{O}\left(\frac{c_1\gamma^{T+t}+c_2\gamma^{(T-t)/2}}{R}\right)$	$\boxed{\mathcal{O}\left(\frac{D}{N^2R}(1+\frac{cD}{N^2R}\mathbb{I}_{T>\hat{T}})\right)}$

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!

Check paper here

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.