Privacy Amplification Strategies towards Better Privacy-Utility Trade-off

Xiang Li

Peking University

April 28, 2023

Outline

Outline

Differential Privacy

Definition ((ε, δ)-DP)

A randomized algorithm $M : \mathbb{X}^n \to \mathbb{O}$ is (ε, δ)-differential private if for every pair of adjacent data sets $D, D' \in \mathbb{X}^n$ that differ by one individual datum and every measurable $S \subseteq \mathbb{O}$,

$$\mathbb{P}\left(M(D) \in S\right) \leq e^{\varepsilon} \cdot \mathbb{P}\left(M(D') \in S\right) + \delta$$

Differential Privacy

Definition ((ε, δ)-DP)

A randomized algorithm $M : \mathbb{X}^n \to \mathbb{O}$ is (ε, δ)-differential private if for every pair of adjacent data sets $D, D' \in \mathbb{X}^n$ that differ by one individual datum and every measurable $S \subseteq \mathbb{O}$,

$$\mathbb{P}\left(M(D) \in S\right) \leq e^{\varepsilon} \cdot \mathbb{P}\left(M(D') \in S\right) + \delta$$

- The probability measure $\mathbb{P}$ is induced by the randomness of M only.
- Many equivalent definitions and other privacy notions.

Differential Privacy

Definition ((ε, δ)-DP)

A randomized algorithm $M : \mathbb{X}^n \to \mathbb{O}$ is (ε, δ)-differential private if for every pair of adjacent data sets $D, D' \in \mathbb{X}^n$ that differ by one individual datum and every measurable $S \subseteq \mathbb{O}$,

$$\mathbb{P}\left(M(D) \in S\right) \leq e^{\varepsilon} \cdot \mathbb{P}\left(M(D') \in S\right) + \delta$$

- The probability measure $\mathbb{P}$ is induced by the randomness of M only.
- Many equivalent definitions and other privacy notions.
- The privacy leakage can be characterized by measuring the dissimilarity between output distributions produced by applying the algorithm to pairs datasets differing in one individual.

Gaussian Mechanism

- Given a dataset D, we wanna compute $f(D)$ whose ℓ_p-norm sensitivity is

$$\Delta_p = \sup_{D \simeq D'} \|f(D) - f(D')\|_p.$$

- A typically mechanism towards (ε, δ)-DP is to add noises

$$M(D) = f(D) + \mathcal{N}_{\text{noise}}.$$

Gaussian Mechanism

- Given a dataset D, we wanna compute $f(D)$ whose ℓ_p-norm sensitivity is

$$\Delta_p = \sup_{D \simeq D'} \|f(D) - f(D')\|_p.$$

- A typically mechanism towards (ε, δ)-DP is to add noises

$$M(D) = f(D) + \mathcal{N}_{\text{noise}}.$$

- Gaussian mechanism [Balle and Wang, 2018] uses $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ with

$$\Phi\left(\frac{\Delta_2}{2\sigma} - \frac{\varepsilon\sigma}{\Delta_2}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta_2}{2\sigma} - \frac{\varepsilon\sigma}{\Delta_2}\right) \leq \delta.$$

- If $\varepsilon, \delta \in (0, 1)$, we can set $\sigma = \Delta_2 \frac{\sqrt{2\log(1.25/\delta)}}{\varepsilon}$.

Outline

Case Study 1: Low-dim Mean Estimation

Given $D = \{x_i\}_{i\in[n]} \subseteq \mathbb{X}^n$ with $x_i \overset{i.i.d.}{\sim} P$, the target is $\mu = \mathbb{E}x$ [Cai et al., 2021].

Theorem

If $\mathcal{P}$ is d-dim σ^2-sub-Gaussian with mean $\|\mu\|_\infty \le \sigma$, for properly small δ and large n, d,

$$\inf_{M\in\mathcal{M}_{\varepsilon,\delta}} \sup_{P\in\mathcal{P}} \mathbb{E}\|M(D) - \mu\|_2^2 \gtrsim \sigma^2 \left(\frac{d}{n} + \frac{d^2\log(1/\delta)}{n^2\varepsilon^2}\right). \tag{1}$$

Case Study 1: Low-dim Mean Estimation

Given $D = \{x_i\}_{i\in[n]} \subseteq \mathbb{X}^n$ with $x_i \overset{i.i.d.}{\sim} P$, the target is $\mu = \mathbb{E}x$ [Cai et al., 2021].

Theorem

If $\mathcal{P}$ is d-dim σ^2-sub-Gaussian with mean $\|\mu\|_\infty \le \sigma$, for properly small δ and large n, d,

$$\inf_{M\in\mathcal{M}_{\varepsilon,\delta}} \sup_{P\in\mathcal{P}} \mathbb{E}\|M(D) - \mu\|_2^2 \gtrsim \sigma^2 \left(\frac{d}{n} + \frac{d^2\log(1/\delta)}{n^2\varepsilon^2}\right). \tag{1}$$

- (1) is achieved (up to log factors) by gaussian noise adding after truncation,

$$M(D) = \frac{1}{n}\sum_{i\in[n]} \Pi_R(x_i) + \mathcal{N}\left(\mathbf{0}, \frac{4R^2 d\log(1/\delta)}{n^2\varepsilon^2}\cdot \mathbf{I}_d\right).$$

where $R \approx \sigma\sqrt{\log n}$.

Case Study 2: Sparse Mean Estimation

Given $D = \{x_i\}_{i\in[n]} \subseteq \mathbb{X}^n$ with $x_i \overset{i.i.d.}{\sim} P$, the target is $\mu = \mathbb{E}x$ [Cai et al., 2021].

Theorem

If $\mathcal{P} = \mathcal{N}(\mu, \sigma^2 \mathbf{I}_d)$ with $\|\mu\|_0 \le s$ and $\|\mu\|_\infty \le 1$,for properly small s, δ and large n, d,

$$\inf_{M\in\mathcal{M}_{\varepsilon,\delta}} \sup_{P\in\mathcal{P}} \mathbb{E}\|M(D) - \mu\|_2^2 \gtrsim \sigma^2 \left(\frac{s \log d}{n} + \frac{s^2 \log^2 d}{n^2 \varepsilon^2} \right). \tag{2}$$

Case Study 2: Sparse Mean Estimation

Given $D = \{x_i\}_{i\in[n]} \subseteq \mathbb{X}^n$ with $x_i \overset{i.i.d.}{\sim} P$, the target is $\mu = \mathbb{E}x$ [Cai et al., 2021].

Theorem

If $\mathcal{P} = \mathcal{N}(\mu, \sigma^2 \mathbf{I}_d)$ with $\|\mu\|_0 \le s$ and $\|\mu\|_\infty \le 1$,for properly small s, δ and large n, d,

$$\inf_{M\in\mathcal{M}_{\varepsilon,\delta}} \sup_{P\in\mathcal{P}} \mathbb{E}\|M(D) - \mu\|_2^2 \gtrsim \sigma^2 \left(\frac{s \log d}{n} + \frac{s^2 \log^2 d}{n^2 \varepsilon^2} \right). \tag{2}$$

- After truncation, (2) is achieved by Laplace noise adding to the privately selected top-s entries (named peeling).

Case Study 3: (Generalized) Linear Regression

Given $D = \{(y_i, x_i)\}_{i\in[n]}$ with $x_i \overset{i.i.d.}{\sim} \mathcal{P}$ and $y_i = x_i^\top \beta + \mathcal{N}(0, \sigma^2)(\|\beta\|_2 \le 1)$, the goal is to estimate β privately [Cai et al., 2021].

Theorem

If $\mathcal{P}$ is d-dim σ^2-sub-Gaussian with $\Sigma_x = \mathrm{Cov}(x_i)$, for properly small δ, ε and large n, d,

$$\inf_{M\in\mathcal{M}_{\varepsilon,\delta}} \sup_{P\in\mathcal{P}} \mathbb{E}\|M(D) - \beta\|_{\Sigma_x}^2 \gtrsim \sigma^2 \left(\frac{d}{n} + \frac{d^2}{n^2\varepsilon^2}\right). \tag{3}$$

Case Study 3: (Generalized) Linear Regression

Given $D = \{(y_i, x_i)\}_{i\in[n]}$ with $x_i \overset{i.i.d.}{\sim} \mathcal{P}$ and $y_i = x_i^\top \beta + \mathcal{N}(0, \sigma^2)(\|\beta\|_2 \le 1)$, the goal is to estimate β privately [Cai et al., 2021].

Theorem

If $\mathcal{P}$ is d-dim σ^2-sub-Gaussian with $\Sigma_x = \mathrm{Cov}(x_i)$, for properly small δ, ε and large n, d,

$$\inf_{M\in\mathcal{M}_{\varepsilon,\delta}} \sup_{P\in\mathcal{P}} \mathbb{E}\|M(D) - \beta\|_{\Sigma_x}^2 \gtrsim \sigma^2 \left(\frac{d}{n} + \frac{d^2}{n^2\varepsilon^2}\right). \tag{3}$$

- (3) is achieved by Differentially Private Linear Regression (DP-GD) where we truncate the gradient and then add gaussian noises.

Case Study 3: (Generalized) Linear Regression

Given $D = \{(y_i, x_i)\}_{i\in[n]}$ with $x_i \overset{i.i.d.}{\sim} \mathcal{P}$ and $y_i = x_i^\top \beta + \mathcal{N}(0, \sigma^2)(\|\beta\|_2 \le 1)$, the goal is to estimate β privately [Cai et al., 2021].

Theorem

If $\mathcal{P}$ is d-dim σ^2-sub-Gaussian with $\Sigma_x = \mathrm{Cov}(x_i)$, for properly small δ, ε and large n, d,

$$\inf_{M\in\mathcal{M}_{\varepsilon,\delta}} \sup_{P\in\mathcal{P}} \mathbb{E}\|M(D) - \beta\|_{\Sigma_x}^2 \gtrsim \sigma^2 \left(\frac{d}{n} + \frac{d^2}{n^2\varepsilon^2}\right). \tag{3}$$

- (3) is achieved by Differentially Private Linear Regression (DP-GD) where we truncate the gradient and then add gaussian noises.
- Similar counterpart results for sparse/generalized linear models [Cai et al., 2020].

Case Study 4: Learning Gaussian Distribution

Theorem (Theorem 1.5 in [Kamath et al., 2019])

Any $(\varepsilon, 0)$-DP algorithm that takes samples from an arbitrary unknown Gaussian distribution $\mathcal{P}$ and outputs a Gaussian distribution $\mathcal{Q}$ such that

$$\mathrm{TV}(\mathcal{P}, \mathcal{Q}) \leq \alpha \text{ with probability} \geq 0.9 \text{ requires } n = \Omega\left(\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon}\right).$$

Case Study 4: Learning Gaussian Distribution

Theorem (Theorem 1.5 in [Kamath et al., 2019])

Any $(\varepsilon, 0)$-DP algorithm that takes samples from an arbitrary unknown Gaussian distribution $\mathcal{P}$ and outputs a Gaussian distribution $\mathcal{Q}$ such that

$$\mathrm{TV}(\mathcal{P}, \mathcal{Q}) \leq \alpha \text{ with probability} \geq 0.9 \text{ requires } n = \Omega\left(\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon}\right).$$

- $\mathrm{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma')) = O(\alpha)$ if $\|\mu - \mu'\|_\Sigma \leq \alpha$ and $\|\Sigma - \Sigma'\|_\Sigma \leq \alpha$.

Case Study 4: Learning Gaussian Distribution

Theorem (Theorem 1.5 in [Kamath et al., 2019])

Any $(\varepsilon, 0)$-DP algorithm that takes samples from an arbitrary unknown Gaussian distribution $\mathcal{P}$ and outputs a Gaussian distribution $\mathcal{Q}$ such that

$$\mathrm{TV}(\mathcal{P}, \mathcal{Q}) \leq \alpha \text{ with probability} \geq 0.9 \text{ requires } n = \Omega\left(\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon}\right).$$

- $\mathrm{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma')) = O(\alpha)$ if $\|\mu - \mu'\|_\Sigma \leq \alpha$ and $\|\Sigma - \Sigma'\|_\Sigma \leq \alpha$.
- Worse by a factor of d due to unknown covariance.

Empirical Risk Minimization (ERM)

- With $D := \{x_1, \cdots, x_n\}$, we want to minimize

$$\mathcal{L}(\theta, D) = \frac{1}{n} \sum_{i \in [n]} \ell(\theta, x_i) \text{ over } \theta \in \mathcal{C} \tag{4}$$

where we assume $\ell(\cdot, x)$ is (i) convex and (ii) 1-Lipschitz in ℓ_p-norm for all x's that is

$$|\ell(\theta, x) - \ell(\theta', x)| \leq \|\theta - \theta'\|_p.$$

Empirical Risk Minimization (ERM)

- With $D := \{x_1, \cdots, x_n\}$, we want to minimize

$$\mathcal{L}(\theta, D) = \frac{1}{n} \sum_{i \in [n]} \ell(\theta, x_i) \text{ over } \theta \in \mathcal{C} \tag{4}$$

  where we assume $\ell(\cdot, x)$ is (i) convex and (ii) 1-Lipschitz in ℓ_p-norm for all x's that is

$$|\ell(\theta, x) - \ell(\theta', x)| \leq \|\theta - \theta'\|_p.$$

- Generalized linear model (GLM) loss: $\ell(\theta, (x, y)) = b(x^\top \theta) - yx^\top \theta$.

Empirical Risk Minimization (ERM)

- With $D := \{x_1, \cdots, x_n\}$, we want to minimize

$$\mathcal{L}(\theta, D) = \frac{1}{n} \sum_{i \in [n]} \ell(\theta, x_i) \text{ over } \theta \in \mathcal{C} \tag{4}$$

where we assume $\ell(\cdot, x)$ is (i) convex and (ii) 1-Lipschitz in ℓ_p-norm for all x's that is

$$|\ell(\theta, x) - \ell(\theta', x)| \leq \|\theta - \theta'\|_p.$$

- Generalized linear model (GLM) loss: $\ell(\theta, (x, y)) = b(x^\top \theta) - yx^\top \theta$.
- The (worst) excess empirical risk of $\hat{\theta} = M(D)$ is

$$(\sup_D) \mathbb{E} \left(\mathcal{L}(\hat{\theta}, D) - \inf_{\theta \in \mathcal{C}} \mathcal{L}(\theta, D) \right). \tag{5}$$

Lower Bounds for ERM

Paper	$\mathcal{C}$-free	p in ℓ_p	Loss	Smooth	Lower bounds
[Bassily et al., 2014]	$\times$	2	GLM	$\times \& \surd$	$\frac{\sqrt{d}}{n\varepsilon}$
[Song et al., 2021]	$\surd$	2	GLM	$\times$	$\frac{\sqrt{\mathrm{rank}}}{n\varepsilon}$
[Asi et al., 2021]	$\surd$	1	general	$\times$	$\frac{\sqrt{d}}{n\varepsilon \log d}$
				$\surd$	$\frac{\sqrt{d}}{n\varepsilon} \wedge \left(\frac{1}{n\varepsilon}\right)^{2/3}$
[Liu and Lu, 2021]	$\surd$	$[1, \infty)$	general	$\times$	$\frac{\sqrt{d \log(1/\delta)}}{n\varepsilon}$

Table: Minimax lower bounds on the excess empirical risk for (ε, δ)-DP convex and Lipschitz ERM.

Lower Bounds for ERM

- The current best upper bound [Bassily et al., 2021] with $\|\mathcal{C}\|_p \le 1$ is

$$\tilde{O}\left(\min\left\{\log d, \frac{1}{p-1}\right\}\frac{\sqrt{d}}{\varepsilon n}\right) \text{ for } p \in [1,2] \text{ and } \tilde{O}\left(\frac{d^{1-1/p}}{\varepsilon n}\right) \text{ for } p \in (2,\infty).$$

- There is a phase transition near $p = 1$ under smoothness assumption.

Lower Bounds for ERM

- The current best upper bound [Bassily et al., 2021] with $\|\mathcal{C}\|_p \leq 1$ is

$$\tilde{O}\left(\min\left\{\log d, \frac{1}{p-1}\right\}\frac{\sqrt{d}}{\varepsilon n}\right) \text{ for } p \in [1,2] \text{ and } \tilde{O}\left(\frac{d^{1-1/p}}{\varepsilon n}\right) \text{ for } p \in (2,\infty).$$

- There is a phase transition near $p = 1$ under smoothness assumption.
- Non-private lower bounds [Agarwal et al., 2009]

p	1	$(1,2]$	$[2,\infty)$
	$\sqrt{\frac{\log d}{n}}$	$\frac{1}{\sqrt{n}}$	$\left(\frac{1}{n}\right)^{1/p} + \frac{d^{1/2-1/p}}{\sqrt{n}}$

Table: Lower bounds for non-private population excess risks

Take-away Messages

- Optimal estimation error = optimal statistical error + cost of privacy
- As long as ε is properly large, privacy could come "for free".

Take-away Messages

- Optimal estimation error = optimal statistical error + cost of privacy
- As long as ε is properly large, privacy could come "for free".
- Lower bounds vary case by case.
- Prior knowledge and problem structures influence the lower bounds.

Outline

Differentially private SGD (DP-SGD [Abadi et al., 2016])

Require: learning rate η_t, noise scale σ, batch size B, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample S_t with sampling probability $q := B/n$

Compute gradient: $\mathbf{g}_t \leftarrow \nabla \mathcal{L}(\theta_t, S_t)$

Clip gradient: $\bar{\mathbf{g}}_t \leftarrow \mathbf{g}_t / \max\left(1, \frac{\|\mathbf{g}_t\|_2}{C}\right)$ (for DP-SGD)

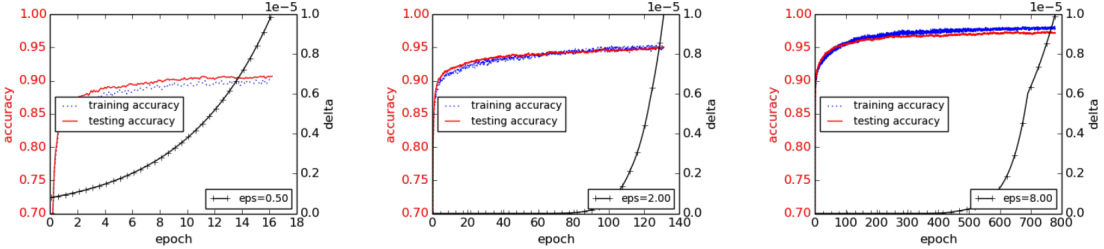
Add Gaussian noise: $\tilde{\mathbf{g}}_t \leftarrow \bar{\mathbf{g}}_t + \mathcal{N}(0, \sigma^2 \boldsymbol{I}_d)$

Descent: $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Projection: $\theta_{t+1} \leftarrow \Pi_{\mathcal{C}}(\theta_{t+1})$ (for Noisy-SGD)

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Numerical Experiments



(1) Large noise (2) Medium noise (3) Small noise

Figure: Accuracy for different noise levels of DP-SGD on MNIST (from [Abadi et al., 2016]).

DP Analysis for DP-SGD

Property (Post-processing)

If M is (ε, δ)-DP, then $F \circ M$ is also (ε, δ)-DP.

Property (Subsampling)

If M is (ε, δ)-DP and U is the Poisson sampler that selects each data with probability q, then $M \circ U$ is $(O(q\varepsilon), q\delta)$-DP.

Property (Advanced composition)

If M is (ε, δ)-DP, then $M^{\circ k}$ is $(O(\varepsilon\sqrt{k\log(1/\delta)}), O(k\delta))$-DP if $\varepsilon \precsim 1/\sqrt{k}$.

DP Analysis for DP-SGD

Property (Post-processing)

If M is (ε, δ)-DP, then $F \circ M$ is also (ε, δ)-DP.

Property (Subsampling)

If M is (ε, δ)-DP and U is the Poisson sampler that selects each data with probability q, then $M \circ U$ is $(O(q\varepsilon), q\delta)$-DP.

Property (Advanced composition)

If M is (ε, δ)-DP, then $M^{\circ k}$ is $(O(\varepsilon\sqrt{k\log(1/\delta)}), O(k\delta))$-DP if $\varepsilon \precsim 1/\sqrt{k}$.

Theorem

For sufficiently small ε, DP-SGD is (ε, δ)-DP if $\sigma \succsim \frac{qC}{\varepsilon}\sqrt{T\log\frac{qT}{\delta}\log\frac{1}{\delta}}$.

Outline

Privacy Amplification Strategies

Strategies that save the privacy budget (ε, δ).

Privacy Amplification Strategies

Strategies that save the privacy budget (ε, δ).

- Subsampling [Balle et al., 2018, 2020]:
  randomize samples used in nested computations.
- Iteration [Feldman et al., 2018]:
  hide intermediate information in iterative optimization.
- Post-processing [Balle et al., 2019]:
  modify the output by stochastic post-processing.
- Data sketching [Yu et al., 2021]:
  add noises embedding a low-dimensional subspace.

Privacy Amplification via Sampling

Definition (Privacy profile)

$$\delta_M(\varepsilon) = \sup_{D \simeq D'} D_{e^\varepsilon}(M(D) \,\|\, M(D'))$$

where D_{e^ε} is the hockey-stick divergence, i.e., f-divergence introduced by $f(t) = (t - e^\varepsilon)_+$.

Privacy Amplification via Sampling

Definition (Privacy profile)

$$\delta_M(\varepsilon) = \sup_{D \simeq D'} D_{e^\varepsilon}(M(D) \,\|\, M(D'))$$

where D_{e^ε} is the hockey-stick divergence, i.e., f-divergence introduced by $f(t) = (t - e^\varepsilon)_+$.

- M is (ε, δ)-DP $\Leftrightarrow \delta_M(\varepsilon) \le \delta$.
- $\delta_M(\varepsilon)$ returns the smallest δ that makes M (ε, δ)-DP for a given ε.

Privacy Amplification via Sampling

Theorem (Theorem 13 in [Balle et al., 2020])

Let U be the Poisson subsampling with probability q, then

$$\delta_{M\circ U}(\varepsilon') \leq q \cdot \delta_M(\varepsilon) \text{ where } \varepsilon' = \log(1 + q(e^{\varepsilon} - 1)).$$

Privacy Amplification via Sampling

Theorem (Theorem 13 in [Balle et al., 2020])

Let U be the Poisson subsampling with probability q, then

$$\delta_{M\circ U}(\varepsilon') \leq q \cdot \delta_M(\varepsilon) \text{ where } \varepsilon' = \log(1 + q(e^{\varepsilon} - 1)).$$

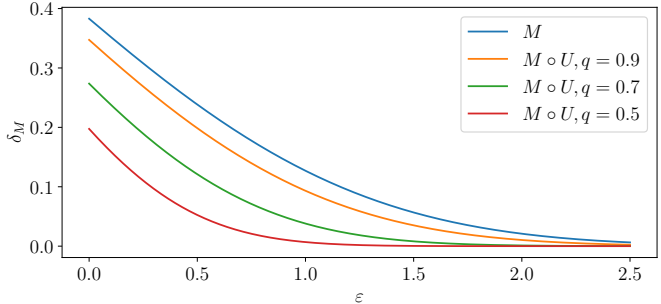


Figure: Privacy profiles of $M(D) = \mathcal{N}(0, 1)$ and $M(D') = \mu + M(D)$ with $\Delta_2 = 1$.

Rényi Differential Privacy: Definition

Definition (Rényi divergence)

Let $1 \leq \alpha < \infty$ and μ, ν be measures with $\mu \ll \nu$. The Renyi divergence of order α between μ and ν is defined as

$$R_\alpha(\mu \parallel \nu) \doteq \frac{1}{\alpha - 1} \ln \int \left(\frac{\mu(z)}{\nu(z)} \right)^\alpha \nu(z) \mathrm{d}z.$$

Rényi Differential Privacy: Definition

Definition (Rényi divergence)

Let $1 \leq \alpha < \infty$ and μ, ν be measures with $\mu \ll \nu$. The Renyi divergence of order α between μ and ν is defined as

$$R_\alpha(\mu \parallel \nu) \doteq \frac{1}{\alpha - 1} \ln \int \left(\frac{\mu(z)}{\nu(z)}\right)^\alpha \nu(z) \mathrm{d}z.$$

Definition ((α, ε)-RDP)

For $1 \leq \alpha < \infty$, a randomized algorithm $M : \mathbb{X}^n \to \mathbb{R}$ is (α, ε)-Rényi differentially private if

$$\sup_{D \backsimeq D'} R_\alpha(M(D) \parallel M(D')) \leq \varepsilon.$$

Rényi Differential Privacy: Properties

Theorem

- *Gaussian mechanism that adds* $\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ *is* $(\alpha, \frac{\alpha \Delta_2^2}{2\sigma^2})$*-RDP.*
- *If* M_1 *is* (α, ε_1)*-RDP and* M_2 *is* (α, ε_2)*-RDP, then* $M_2 \circ M_1$ *is* $(\alpha, \varepsilon_1 + \varepsilon_2)$*-RDP.*
- *If* M *is* (α, ε)*-RDP, then it is* $\left(\varepsilon + \frac{\ln(1/\delta)}{\alpha - 1}, \delta\right)$*-DP for any* $0 < \delta < 1$.

Privacy Amplification via Iteration: Key Result

Theorem (Theorem 1 in [Feldman et al., 2018])

Let $x_0 \in \mathbb{R}^d$ and let X_T be obtained from x_0 by iterating

$$x_{t+1} \doteq \psi_{t+1}(x_t) + Z_{t+1} \text{ with } Z_{t+1} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$$

for some sequence of contractive (or 1-Lipschitz) maps $\{\psi_t\}_{t=1}^T$. Let X'_T denote the output of the same process started at some x'_0. Then for every $\alpha \geq 1$,

$$R_\alpha(X_T \parallel X'_T) \leq \frac{\alpha \|x_0 - x'_0\|^2}{2T\sigma^2}.$$

Privacy Amplification via Iteration: Key Result

Theorem (Theorem 1 in [Feldman et al., 2018])

Let $x_0 \in \mathbb{R}^d$ and let X_T be obtained from x_0 by iterating

$$x_{t+1} \doteq \psi_{t+1}(x_t) + Z_{t+1} \text{ with } Z_{t+1} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$$

for some sequence of contractive (or 1-Lipschitz) maps $\{\psi_t\}_{t=1}^T$. Let X'_T denote the output of the same process started at some x'_0. Then for every $\alpha \geq 1$,

$$R_\alpha(X_T \parallel X'_T) \leq \frac{\alpha \|x_0 - x'_0\|^2}{2T\sigma^2}.$$

Example

If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and β-smooth, then $\phi(x) \doteq x - \eta \nabla f(x)$ is contractive if $\eta \leq 2/\beta$.

Analysis for Noisy-SGD with One-pass Shuffling

Require: Data set $D = \{x_1, \ldots, x_n\}$, learning rate η, starting point $\theta_0 \in \mathcal{C}$, noise scale σ.

for $t \in \{0, \ldots, n-1\}$ **do**

$v_{t+1} \leftarrow w_t - \eta(\nabla_\theta \ell(\theta_t, x_{t+1}) + \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d))$

$\theta_{t+1} \leftarrow \Pi_{\mathcal{C}}(v_{t+1})$

return the final iterate θ_n or random stopping iterate θ_T where $T \sim \mathrm{Uniform}([n])$

Analysis for Noisy-SGD with One-pass Shuffling

Require: Data set $D = \{x_1, \dots, x_n\}$, learning rate η, starting point $\theta_0 \in \mathcal{C}$, noise scale σ.
  for $t \in \{0, \dots, n-1\}$ **do**
    $v_{t+1} \leftarrow w_t - \eta(\nabla_\theta \ell(\theta_t, x_{t+1}) + \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d))$
    $\theta_{t+1} \leftarrow \Pi_{\mathcal{C}}(v_{t+1})$
  return the final iterate θ_n or random stopping iterate θ_T where $T \sim \mathrm{Uniform}([n])$

Theorem (Theorem 23 and 26 in [Feldman et al., 2018])

Recall that $\ell(\cdot, x)$ is convex L-Lipschitz and β-smooth functions over the convex set $\mathcal{C} \subseteq \mathbb{R}^d$. Assume $\eta \le 2/\beta, \alpha > 1$, starting point $w_0 \in \mathcal{C}$, $\sigma \ge L\sqrt{2(\alpha-1)\alpha}$, and dataset $S \in \mathbb{X}^n$. For the above algorithm,

- *The last iterate $\theta_n(D)$ is $\left(\alpha, \frac{\alpha L^2}{(n+1-j)\sigma^2}\right)$-RDP for the j-th data.*
- *The random-stop iterate $\theta_T(D)$ is $\left(\alpha, \frac{4\alpha L^2 \cdot \ln n}{n\sigma^2}\right)$-RDP.*

Analysis for Noisy-SGD with One-pass Shuffling: Limitation

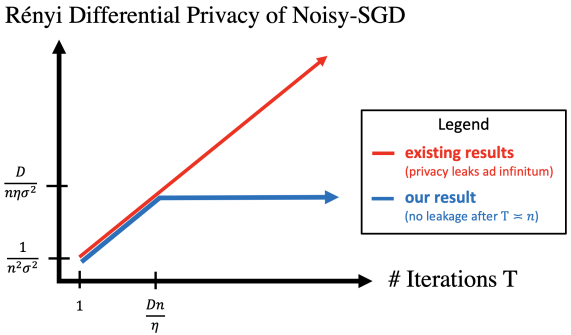
- Only one pass of data $\Longrightarrow$ Use the composition theorem to analysis multiple passes $\Longrightarrow$ unbounded RDP loss.
- The often use of batched samples.
- Unclear for strongly convex case.

Analysis for Noisy-SGD with One-pass Shuffling: Limitation

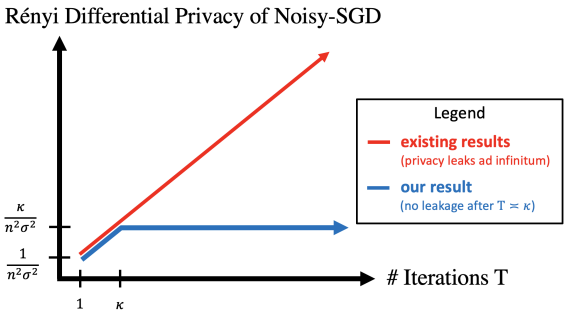
- Only one pass of data $\Longrightarrow$ Use the composition theorem to analysis multiple passes $\Longrightarrow$ unbounded RDP loss.
- The often use of batched samples.
- Unclear for strongly convex case.

All the above addressed by Altschuler and Talwar [2022] recently.

Analysis for Noisy-SGD with Hidden States



(a) Convex and smooth losses.



(b) Strongly convex and smooth losses.

Figure: The RDP loss of Noisy-SGD is identical except for an unavoidable worsening in an initial phase (cited from [Altschuler and Talwar, 2022]).

Privacy Amplification via Post-processing

Definition (Markov operator)

Let $K : \mathbb{O} \to \mathcal{P}(\mathbb{O})$ be a Markov operator with $x \mapsto K(x)(E)$ a measure on $\mathbb{O}$.

- *For a distribution $\nu \in \mathbb{P}(\mathbb{O})$, $(\nu K)(E) = \int K(x)(E)\nu(\mathrm{d}x)$.*
- *For a function $f : \mathbb{O} \to \mathbb{R}$, $(Kf)(x) = \int f(y)K(x, \mathrm{d}y)$.*

Privacy Amplification via Post-processing

Definition (Markov operator)

Let $K : \mathbb{O} \to \mathcal{P}(\mathbb{O})$ be a Markov operator with $x \mapsto K(x)(E)$ a measure on $\mathbb{O}$.

- *For a distribution $\nu \in \mathbb{P}(\mathbb{O})$, $(\nu K)(E) = \int K(x)(E)\nu(\mathrm{d}x)$.*
- *For a function $f : \mathbb{O} \to \mathbb{R}$, $(Kf)(x) = \int f(y)K(x, \mathrm{d}y)$.*

Theorem (Theorem 1 in [Balle et al., 2019])

If the Markov operator K satisfies $\sup_{x,x'} \mathrm{TV}(K(x), K(x')) \leq \gamma \leq 1$, then for any (ε, δ)-DP mechanism M, $K \circ M$ is $(\varepsilon, \gamma\delta)$-DP.

Diffusion Mechanism

- Consider a family of Markov operators $\mathbf{P} = \{P_t\}_{t\geq 0}$ satisfying $P_t P_s = P_{t+s}$.
- Assume (i) there exists a unique non-negative invariant measure λ, (ii) P_t has a symmetric kernel $p_t(x,y) = p_t(y,x)$ w.r.t. λ, and (iii) its generator L defined by $Lf = \frac{\mathrm{d}}{\mathrm{d}t}(P_t f)|_{t=0}$ satisfies $L\phi(f) = \phi'(f)Lf + \phi''(f)\Gamma(f)$ for any differentiable $\phi : \mathbb{R} \to \mathbb{R}$.
- Γ is the *carré du champ* operator: $\Gamma(f,g) = \frac{1}{2}(L(fg) - fLg - gLf)$ and $\Gamma(f) = \Gamma(f,f)$.
- L is a second-order differential operator without constant terms.

Diffusion Mechanism

- Consider a family of Markov operators $\mathbf{P} = \{P_t\}_{t\geq 0}$ satisfying $P_t P_s = P_{t+s}$.
- Assume (i) there exists a unique non-negative invariant measure λ, (ii) P_t has a symmetric kernel $p_t(x,y) = p_t(y,x)$ w.r.t. λ, and (iii) its generator L defined by $Lf = \frac{\mathrm{d}}{\mathrm{d}t}(P_t f)|_{t=0}$ satisfies $L\phi(f) = \phi'(f)Lf + \phi''(f)\Gamma(f)$ for any differentiable $\phi : \mathbb{R} \to \mathbb{R}$.
- Γ is the *carré du champ* operator: $\Gamma(f,g) = \frac{1}{2}(L(fg) - fLg - gLf)$ and $\Gamma(f) = \Gamma(f,f)$.
- L is a second-order differential operator without constant terms.

Theorem (Theorem 6 in [Balle et al., 2019])

The diffusion mechanism $M_t^f(D) = P_t(f(D))$ is $(\alpha, \alpha\Lambda(t))$-RDP for any $\alpha > 1$ and $t > 0$ where

$$\Lambda(t) = \sup_{D \simeq D'} \int_t^\infty \sup_y \Gamma\left(\log \frac{p_s(f(D), y)}{p_s(f(D'), y)}\right) \mathrm{d}s.$$

Ornstein-Uhlenbeck Mechanism

- Ornstein-Uhlenbeck process: $\mathrm{d}X_t = -\theta X_t \mathrm{d}t + \sqrt{2}\rho \mathrm{d}W_t$
- $P_t(x) = \mathcal{N}\left(e^{-\theta t}x, \frac{\rho^2}{\theta}\left(1 - e^{-2\theta t}\right)\mathbf{I}_d\right)$ and $\lambda = \mathcal{N}(0, \rho^2/\theta \mathbf{I}_d)$.
- $Lf = -\theta\langle x, \nabla f\rangle + \rho^2 \nabla^2 f$ and $\Gamma(f, g) = \rho^2\langle \nabla f, \nabla g\rangle$.
- Ornstein-Uhlenbeck (OU) mechanism:

$$M_t^f(D) = e^{-\theta t} f(D) + \mathcal{N}\left(0, \frac{\rho^2}{\theta}\left(1 - e^{-2\theta t}\right)\mathbf{I}_d\right).$$

Ornstein-Uhlenbeck Mechanism

- Ornstein-Uhlenbeck process: $\mathrm{d}X_t = -\theta X_t \mathrm{d}t + \sqrt{2}\rho \mathrm{d}W_t$
- $P_t(x) = \mathcal{N}\left(e^{-\theta t}x, \frac{\rho^2}{\theta}\left(1 - e^{-2\theta t}\right)\mathbf{I}_d\right)$ and $\lambda = \mathcal{N}(0, \rho^2/\theta \mathbf{I}_d)$.
- $Lf = -\theta\langle x, \nabla f\rangle + \rho^2 \nabla^2 f$ and $\Gamma(f, g) = \rho^2\langle \nabla f, \nabla g\rangle$.
- Ornstein-Uhlenbeck (OU) mechanism:

$$M_t^f(D) = e^{-\theta t} f(D) + \mathcal{N}\left(0, \frac{\rho^2}{\theta}\left(1 - e^{-2\theta t}\right)\mathbf{I}_d\right).$$

Corollary (Corollary 7 in [Balle et al., 2019])

The OU mechanism satisfies $(\alpha, \alpha\Gamma(t))$*-RDP with*

$$\Gamma(t) = \frac{\theta \Delta_2^2}{2\rho^2(e^{2\theta t} - 1)}.$$

Privacy Amplification via Data Sketching

- Add noises to gradients: $\tilde{\mathbf{g}} \leftarrow \bar{\mathbf{g}} + \mathbf{z}$ with $\mathbf{z} = \mathcal{N}(0, \sigma^2 \boldsymbol{I}_d)$.
- $\mathbb{E}\|\mathbf{z}\|_2^2$ scales linearly with the model dimension d.
- In DL, gradients live on a very low dimensional manifold.

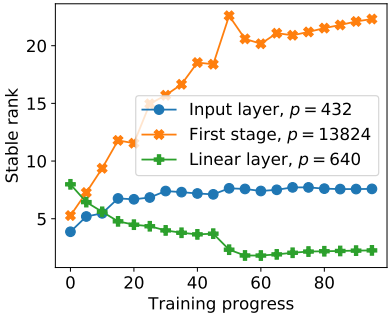


Figure: Stable rank ($\|A\|_F^2/\|A\|_2^2$) of ResNet20 on CIFAR-10.

Privacy Amplification via Data Sketching

- Add noises to gradients: $\tilde{\mathbf{g}} \leftarrow \bar{\mathbf{g}} + \mathbf{z}$ with $\mathbf{z} = \mathcal{N}(0, \sigma^2 \boldsymbol{I}_d)$.
- $\mathbb{E}\|\mathbf{z}\|_2^2$ scales linearly with the model dimension d.
- In DL, gradients live on a very low dimensional manifold.

Gradient Embedding Perturbation:

1. Compute the anchor subspace $\mathbf{B}$ from a public dataset.
2. Project the private gradients $\mathbf{G}$ into the subspace $\mathbf{B}$.
3. Obtain $\mathbf{W} = \mathbf{G}\mathbf{B}^\top$ and residual $\mathbf{R} = \mathbf{G} - \mathbf{W}\mathbf{B}$.
4. Perturb them and obtain $\widetilde{\mathbf{w}}$ and $\widetilde{\mathbf{r}}$.
5. GEP releases $\widetilde{\mathbf{w}}^\top \mathbf{B} + \widetilde{\mathbf{r}}$, while B-GEP releases $\widetilde{\mathbf{w}}^\top \mathbf{B}$ only.

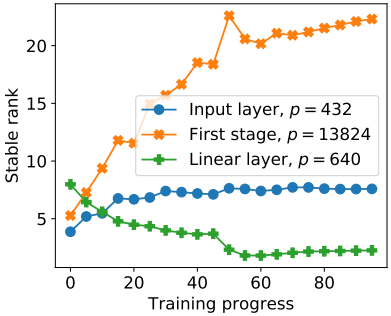


Figure: Stable rank ($\|A\|_F^2/\|A\|_2^2$) of ResNet20 on CIFAR-10.

Privacy Amplification via Data Sketching

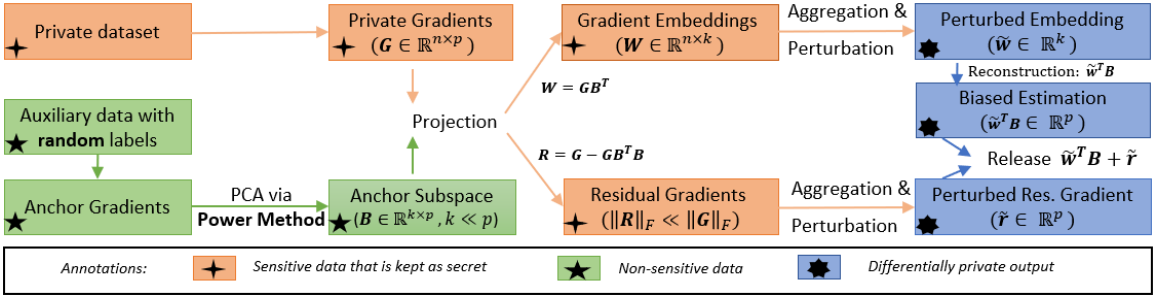


Figure: Overview of Gradient Embedding Perturbation (GEP) algorithm [Yu et al., 2021].

Privacy Amplification via Data Sketching: Results

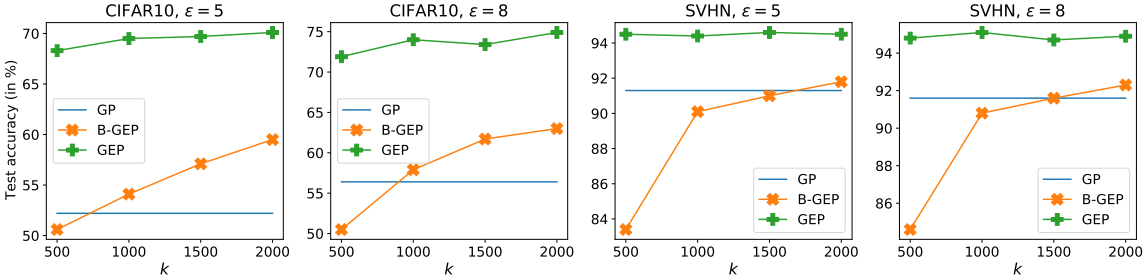


Figure: Test accuracy when varying the dimension of anchor subspace.

Take-away Messages

Take-away Messages

- Privacy amplification via subsampling lowers the whole privacy profile δ_M.

Take-away Messages

- Privacy amplification via subsampling lowers the whole privacy profile δ_M.
- Privacy amplification via iteration assumes secrecy of intermediate computations and rules out sharing intermediate updates (a standard step in federated learning).
- Given a reasonable privacy budget (ε, δ), no limit for the number of iterations that Noisy-SGD can be run.

Take-away Messages

- Privacy amplification via subsampling lowers the whole privacy profile δ_M.
- Privacy amplification via iteration assumes secrecy of intermediate computations and rules out sharing intermediate updates (a standard step in federated learning).
- Given a reasonable privacy budget (ε, δ), no limit for the number of iterations that Noisy-SGD can be run.
- Privacy amplification via post-processing perturbs outputs; privacy loss $\downarrow$ but bias $\uparrow$.

Take-away Messages

- Privacy amplification via subsampling lowers the whole privacy profile δ_M.
- Privacy amplification via iteration assumes secrecy of intermediate computations and rules out sharing intermediate updates (a standard step in federated learning).
- Given a reasonable privacy budget (ε, δ), no limit for the number of iterations that Noisy-SGD can be run.
- Privacy amplification via post-processing perturbs outputs; privacy loss $\downarrow$ but bias $\uparrow$.
- Privacy amplification via sketching relies on low-dimensional projection and has better empirical performance.

Take-away Messages

- Privacy amplification via subsampling lowers the whole privacy profile δ_M.
- Privacy amplification via iteration assumes secrecy of intermediate computations and rules out sharing intermediate updates (a standard step in federated learning).
- Given a reasonable privacy budget (ε, δ), no limit for the number of iterations that Noisy-SGD can be run.
- Privacy amplification via post-processing perturbs outputs; privacy loss $\downarrow$ but bias $\uparrow$.
- Privacy amplification via sketching relies on low-dimensional projection and has better empirical performance.
- To ease privacy analysis for a particular problem/algorithm, it's important to select the appropriate privacy notion.

Outline

Summary

Summary

- Several lower bounds for the privacy-utility trade-off.

Summary

- Several lower bounds for the privacy-utility trade-off.
- Several privacy amplification strategies to improve the privacy-utility trade-off.

Summary

- Several lower bounds for the privacy-utility trade-off.
- Several privacy amplification strategies to improve the privacy-utility trade-off.

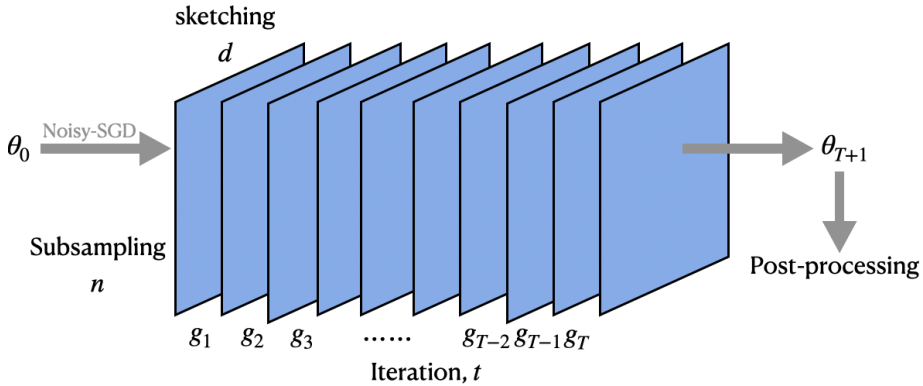


Figure: Summary of four privacy amplification strategies.

References I

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

Jason Altschuler and Kunal Talwar. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. In *Advances in Neural Information Processing Systems*, 2022.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. In *International Conference on Machine Learning*, pages 393–403. PMLR, 2021.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR, 2018.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Borja Balle, Gilles Barthe, Marco Gaboardi, and Joseph Geumlek. Privacy amplification by mixing and diffusion mechanisms. In *Advances in neural information processing systems*, volume 32, 2019.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy profiles and amplification by subsampling. *Journal of Privacy and Confidentiality*, 10(1), 2020.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE, 2014.

Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pages 474–499. PMLR, 2021.

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. *arXiv preprint arXiv:2011.03900*, 2020.

References II

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *Annual Symposium on Foundations of Computer Science*, pages 521–532. IEEE, 2018.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.

Daogao Liu and Zhou Lu. Lower bounds for differentially private erm: Unconstrained and non-euclidean. *arXiv preprint arXiv:2105.13637*, 2021.

Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private GLMs. In *International Conference on Artificial Intelligence and Statistics*, pages 2638–2646. PMLR, 2021.

Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *International Conference on Learning Representations*, 2021.