Lecture 14 Data-Dependent DP Algorithm design

Yu-Xiang Wang

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

Recap: Differentially Private Machine Learning

- Private learning from a finite class is easy
- Private learning from an infinite class is hard (in general)
- Let's restrict our attention to the Lipschitz losses

Recap: Convex empirical risk minimization

- Posterior sampling (i.e., exponential mechanism)

- Output perturbation / Objective perturbation

- NoisyGD and NoisySGD

Recap: NoisyGD summary

	Lipschitz + convex	Lipschitz + Smooth + convex	Smooth + Lipschitz + convex + GLM
Output Pert	$\frac{d^{1/4}L\|\theta^*\|\log(\frac{1}{\delta})^{1/4}}{n^{1/2}\epsilon^{1/2}}$	$\frac{d^{1/3}\beta^{1/3}L^{2/3}\|\theta^*\|^{4/3}\log(\frac{1}{\delta})^{1/3}}{n^{2/3}\epsilon^{2/3}}$	Same as left
ObjPert	Not applicable	$\frac{dL\|\theta^*\|\sqrt{\log(\frac{1}{\delta})}}{n\epsilon}$ Lower order terms and dependence on β hidden.	$\frac{\sqrt{d}L\|\theta^*\|\sqrt{\log(\frac{1}{\delta})}}{n\epsilon}$
NoisyGD	$\frac{\sqrt{d}L\|\theta^*\|\sqrt{\log(\frac{1}{\delta})}}{n\epsilon}$	$\frac{\sqrt{d}L\|\theta^*\|\sqrt{\log(\frac{1}{\delta})}}{n\epsilon}$	$\frac{\sqrt{d}L\|\theta^*\|\sqrt{\log(\frac{1}{\delta})}}{n\epsilon}$

	Lipschitz + Strongly convex	Lipschitz + Smooth + Nonconvex
NoisyGD	$\frac{dL^2\log(1/\delta)}{n\lambda\epsilon^2}$	$\frac{\sqrt{n\beta dL^2(f(\theta_1)-f^*)\log(1/\delta)}}{n\epsilon}$ Stationary point convergence

Recap: Comparing NoisyGD and NoisySGD computationally

- Both optimal information-theoretically.
  - If we ignore computation and add very large noise, but use infinitesimal step-size
- Table to compare computation
  - in terms of the number of incremental gradient calls to achieve information theoretic limit up to a constant

	Lipschitz + Smooth + Convex	Lipschitz + Convex	Lipschitz + Strongly convex
NoisyGD	$\frac{n^2\beta\|x_1 - x^*\|\sqrt{\rho}}{\sqrt{d}L}$	$\frac{n^3\rho}{d}$	$\frac{n^3\rho}{\lambda}$
NoisySGD	$\frac{n^{3/2}\beta^{1/2}\|x_1 - x^*\|\rho^{1/2}}{d^{1/4}L^{1/2}} + \frac{n^2\rho}{d}$	$\frac{n^2\rho^{3/4}}{d^{1/2}} + \frac{n^2\rho}{d}$	$\frac{n^2\rho^{3/4}}{d^{1/2}} + \frac{n^2\rho}{d}$

Open problem: what is the optimal computational complexity?

Recap: Deep Learning with DP

- NoisySGD with per-example gradient clipping
  - The only practical / popular algorithm
  - Empirical research questions: What are tricks to improve NoisySGD?
  - Theoretical open problem: what exactly is the effect of gradient clipping in training? How does it work?

- Assume access to some (unlabeled) public data
  - Private Aggregation of Teacher Ensembles.
  - PrivateKNN

Very few public data points are needed in PATE… also it learns all VC-classes in the realizable setting.

Table 1: Summary of our results: excess risk bounds for PATE algorithms.

Algorithm	**PATE** (Gaussian Mech.) Papernot et al. [2017]	**PATE** (SVT-based) Bassily et al. [2018a]	**PATE** (SVT-based) This paper	**PATE** (Active Learning) This paper
Realizable	$\tilde{O}\left(\frac{d}{(n\epsilon)^{2/3}} \vee \frac{d}{m}\right)$	$\tilde{O}\left(\frac{d}{(n\epsilon)^{2/3}} \vee \sqrt{\frac{d}{m}}\right)$	$\tilde{O}\left(\frac{d^{3/2}}{n\epsilon} \vee \frac{d}{m}\right)$	$\tilde{O}\left(\frac{d^{3/2}\theta^{1/2}}{n\epsilon} \vee \frac{d}{m}\right)$
τ-TNC	$\tilde{O}\left(\left(\frac{d^{3/2}}{n\epsilon}\right)^{\frac{2\tau}{4-\tau}} \vee \frac{d}{m}\right)$	same as agnostic	$\tilde{O}\left(\left(\frac{d^{3/2}}{n\epsilon}\right)^{\frac{\tau}{2-\tau}} \vee \frac{d}{m}\right)$	$\tilde{O}\left(\left(\frac{d^{3/2}\theta^{1/2}}{n\epsilon}\right)^{\frac{\tau}{2-\tau}} \vee \frac{d}{m}\right)$
Agnostic (vs h^*)	$\Omega(\texttt{Err}(h^*))$ required.	$13\texttt{Err}(h^*)+$ $\tilde{O}\left(\frac{d^{3/5}}{n^{2/5}\epsilon^{2/5}} \vee \sqrt{\frac{d}{m}}\right)$	$\Omega(\texttt{Err}(h^*))$ required.	$\Omega(\texttt{Err}(h^*))$ required.
Agnostic (vs $h_\infty^{\texttt{agg}}$)	-	-	Consistent under weaker conditions.	-

Liu, Zhu, Chaudhuri and W. (2020) "Revisiting model-agnostic private learning". AISTATS and JMLR.
https://arxiv.org/pdf/2011.03186.pdf

This lecture

- Going beyond the worst case!

- Smoothed Sensitivity and the Median

- Concentrated DP of Smoothed Sensitivity-based algorithm

Reading materials

- Nissim, Raskhodnikova, Smith 2011 “Smooth Sensitivity and Sampling in Privacy Data Analysis”: https://cs-people.bu.edu/ads22/pubs/NRS07/NRS07-full-draft-v1.pdf

- Bun and Steinke 2019: “Average cases averages” https://arxiv.org/abs/1906.02830

Recap: Private query release

- For example, linear queries

- Laplace mechanism / Gaussian mechanism

- Global sensitivity

An example when the global sensitivity approach is very inefficient

- Median query:

- Example:

Another example: linear regression

- The output perturbation mechanism, revisited

- The global sensitivity approach does not exploit the fact that the input dataset is well-conditioned

Local Sensitivity measures the stability of a query at a particular input dataset.

$$\mathrm{LS}_q(x) = \max \left\{ q(x) - q(x')| : x' \sim x \right\}.$$

- Example: median

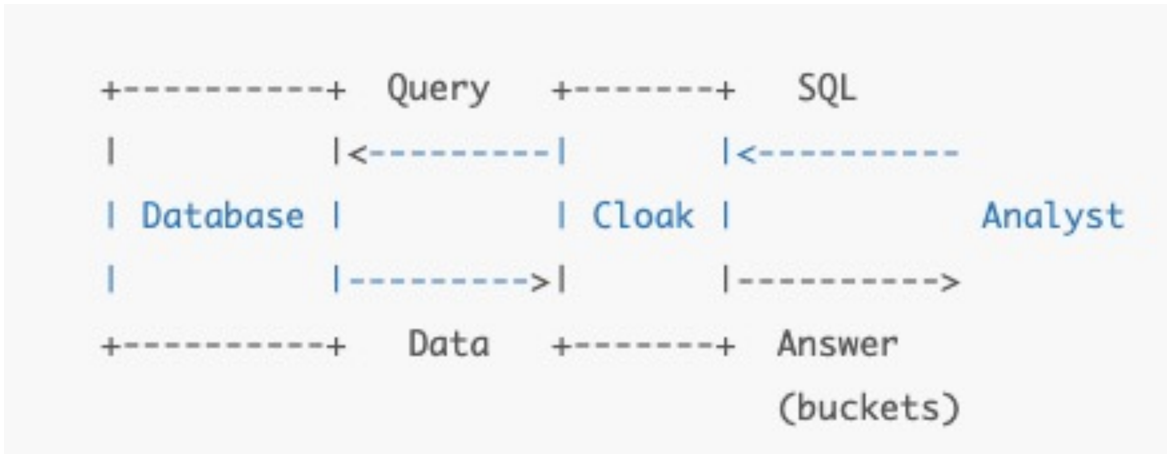
- Example: linear regression

The issue of calibrating noise to local sensitivity

- Example of the median

- In conclusion: the magnitude of the noise may reveal sensitive information!

Diffix and “Sticky Noise”



Implementing a bunch of heuristics to protect against known attacks.
Decide how much noise to add by the specific dataset and how sensitive the question is.

From Creater of Diffix:
“anonymizing SQL interface [that] sits in front of your data and enables you to conduct ad hoc analytics — fully privacy preserving and GDPR-compliant.”

Attack on Diffix

When the Signal is in the Noise:
Exploiting Diffix's Sticky Noise

Andrea Gadotti*[a], Florimond Houssiau*[a], Luc Rocher*[a,b], Benjamin Livshits[a], and Yves-Alexandre de Montjoye†[a]

[a]*Department of Computing and Data Science Institute, Imperial College London*
[b]*ICTEAM, Université catholique de Louvain*

Link to the paper:
https://www.usenix.org/system/files/sec19fall_gadotti_prepub.pdf

Also see this nice post: https://differentialprivacy.org/diffix-attack/

“Data-dependent DP mechanism” aims at more stably calibrating noise to local sensitivity (at least for query releases)

- A number of different approaches:
  - Smooth sensitivity
  - Propose-test-release
  - Privately bounding the local-sensitivity
  - Stability-based query release (Distance2Stability)

Smooth Sensitivity

Definition 2.2 (Smooth sensitivity). *For $\beta > 0$, the β-smooth sensitivity of f is*

$$S^*_{f,\beta}(x) = \max_{y \in D^n} \left(LS_f(y) \cdot e^{-\beta d(x,y)} \right).$$

- Illustration

Smooth sensitivity satisfies a smoothing property, and it is the optimal bound satisfying this property

- Two properties that one should satisfy to smooth out the local sensitivity

$$\forall x \in D^n: \qquad S(x) \geq LS_f(x) \;;$$
$$\forall x, y \in D^n, d(x,y) = 1: \qquad S(x) \leq e^{\beta} S(y) \;.$$

- Smooth sensitivity is the optimal bound

LEMMA 2.3. *$S^*_{f,\beta}$ is a β-smooth upper bound on LS_f. In addition, $S^*_{f,\beta}(x) \leq S(x)$ for all $x \in D^n$ for every β-smooth upper bound S on LS_f.*

What noise to add?

Notation. For a subset $\mathcal{S}$ of $\mathbb{R}^d$, we write $\mathcal{S}+\Delta$ for the set $\{z+\Delta \mid z \in \mathcal{S}\}$, and $e^\lambda \cdot \mathcal{S}$ for the set $\{e^\lambda \cdot z \mid z \in \mathcal{S}\}$. We also write $a \pm b$ for the interval $[a-b, a+b]$.

Definition 2.5 (Admissible Noise Distribution)**.** *A probability distribution on $\mathbb{R}^d$, given by a density function h, is (α,β)-*admissible *(with respect to ℓ_1) if, for $\alpha=\alpha(\epsilon,\delta)$, $\beta=\beta(\epsilon,\delta)$, the following two conditions hold for all $\Delta \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$ satisfying $\|\Delta\|_1 \le \alpha$ and $|\lambda| \le \beta$, and for all measurable subsets $\mathcal{S} \subseteq \mathbb{R}^d$:*

Sliding Property: $$\Pr_{Z\sim h}\Big[Z \in \mathcal{S}\Big] \le e^{\frac{\epsilon}{2}} \cdot \Pr_{Z\sim h}\Big[Z \in \mathcal{S}+\Delta\Big] + \frac{\delta}{2}.$$

Dilation Property: $$\Pr_{Z\sim h}\Big[Z \in \mathcal{S}\Big] \le e^{\frac{\epsilon}{2}} \cdot \Pr_{Z\sim h}\Big[Z \in e^\lambda \cdot \mathcal{S}\Big] + \frac{\delta}{2}.$$

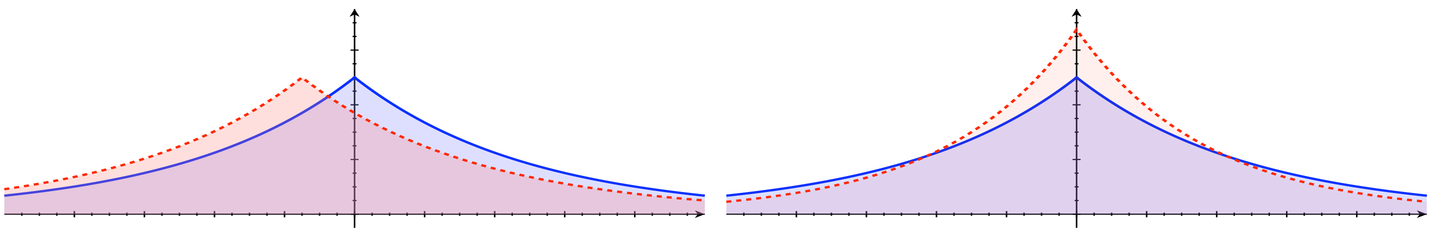


Figure 1: Sliding and dilation for the Laplace distribution with p.d.f. $h(z)=\frac{1}{2}e^{-|z|}$, plotted as a solid line. The dotted lines plot the densities $h(z+0.3)$ (left) and $e^{0.3}h(e^{0.3}z)$ (right).

- Then $\mathcal{A}(x) = f(x) + \frac{S(x)}{\alpha} \cdot Z$ satisfies (ε,δ)-DP.

Privacy analysis

Example: Cauchy-Noise, Laplace-noise, Gaussian noise

Lemma 2.7. *For any $\gamma > 1$, the distribution with density $h(z) \propto \frac{1}{1+|z|^\gamma}$ is $(\frac{\epsilon}{2(\gamma+1)}, \frac{\epsilon}{2(\gamma+1)})$-admissible (with $\delta = 0$). Moreover, the d-dimensional product of independent copies of h is $(\frac{\epsilon}{2(\gamma+1)}, \frac{\epsilon}{2d(\gamma+1)})$- admissible.*

Lemma 2.9. *For $\epsilon, \delta \in (0,1)$, the d-dimensional Laplace distribution, $h(z) = \frac{1}{2^d} \cdot e^{-\|z\|_1}$, is (α, β)-admissible with $\alpha = \frac{\epsilon}{2}$, and $\beta = \frac{\epsilon}{2\rho_{\delta/2}(\|Z\|_1)}$, where $Z \sim h$. In particular, it suffices to use $\alpha = \frac{\epsilon}{2}$ and $\beta = \frac{\epsilon}{4(d+\ln(2/\delta))}$. For $d = 1$, it suffices to use $\beta = \frac{\epsilon}{2\ln(2/\delta)}$.*

Lemma 2.10 (Gaussian Distribution)**.** *For $\epsilon, \delta \in (0,1)$, the d-dimensional Gaussian distribution, $h(z) = \frac{1}{(2\pi)^{d/2}} \cdot e^{-\frac{1}{2}\|z\|_2^2}$, is (α, β)-admissible* for the Euclidean metric *with $\alpha = \frac{\epsilon}{5\rho_{\delta/2}(Z_1)}$, and $\beta = \frac{\epsilon}{2\rho_{\delta/2}(\|Z\|_2^2)}$, where $Z = (Z_1, ..., Z_d) \sim h$.*

In particular, it suffices to take $\alpha = \frac{\epsilon}{5\sqrt{2\ln(2/\delta)}}$ and $\beta = \frac{\epsilon}{4(d+\ln(2/\delta))}$.

How to compute smooth sensitivity (or an upper bound of it?)

DEFINITION 2.2 (SMOOTH SENSITIVITY). *For $\beta > 0$, the β-smooth sensitivity of f is*

$$S^*_{f,\beta}(x) = \max_{y \in D^n} \left(LS_f(y) \cdot e^{-\beta d(x,y)} \right).$$

- An easier way to solve this optimization

$$S^*_{f,\epsilon}(x) \quad = \quad \max_{k=0,1,\ldots,n} e^{-k\epsilon} \left(\max_{y:\ d(x,y)=k} LS_f(y) \right)$$

Example: smooth sensitivity of median

- Recall: $$0 \leq x_1 \leq \cdots \leq x_n \leq \Lambda.$$

$$GS_{f_{med}} = \Lambda \qquad LS_{f_{med}} = \max(x_m - x_{m-1}, x_{m+1} - x_m) \text{ for } m = \frac{n+1}{2}$$

- Now: $$S^*_{f,\epsilon}(x) \;=\; \max_{k=0,1,\ldots,n} e^{-k\epsilon} \left(\max_{y:\ d(x,y)=k} LS_f(y) \right)$$

$$\max_{y:\ d(x,y)\leq k} LS(y) = \max_{0\leq t\leq k+1} \left(x_{m+t} - x_{m+t-k-1}\right) .$$

Checkpoint: smooth sensitivity

- We cannot calibrate noise to local sensitivity
  - Because noise-level itself may be sensitive
- Idea: construct smooth upper bound of local sensitivity
- Noise that satisfies stability under “translation” and “scaling” are admissible

Concentrated DP analysis of Smoothed Sensitivity

- Adding log-normal noise

$$Z = X \cdot e^{\sigma Y}$$

- X drawn from Laplace and Y from a standard Normal.

Proposition 3. *Let* $f : \mathcal{X}^n \to \mathbb{R}$ *and let* $Z \leftarrow \mathsf{LLN}(\sigma)$ *for some* $\sigma > 0$*. Then, for all* $s, t > 0$*, the algorithm* $M(x) = f(x) + \frac{1}{s} \cdot \mathsf{S}_f^t(x) \cdot Z$ *guarantees* $\frac{1}{2}\varepsilon^2$*-CDP for* $\varepsilon = t/\sigma + e^{3\sigma^2/2}s$*.*

Bun and Steinke (2019): "Average case averages": https://arxiv.org/pdf/1906.02830.pdf

Summary of the noises that are known to work

- Cauchy distribution
- Student t-distribution

- Laplace-log-normal
- Uniform-log-normal
- Arcsinh-normal

- Gaussian
- Laplace

Sketch of the proof for the Laplace-Log-Normal

- Let's say for all neighboring datasets

$$|f(x) - f(x')| \leq g(x) \qquad \text{and} \qquad e^{-t} g(x) \leq g(x') \leq e^t g(x).$$

- Algorithm: $M(x) = f(x) + \frac{g(x)}{s} \cdot Z \qquad \text{for} \qquad Z \leftarrow \mathsf{LLN}(\sigma).$

- We have that $\mathrm{D}_\alpha\left(M(x) \| M(x')\right) = \mathrm{D}_\alpha\left(Z \middle\| \frac{f(x') - f(x)}{g(x)} \cdot s + \frac{g(x')}{g(x)} \cdot Z\right).$

Technical tools

- Group privacy for CDP:

Lemma 11. *Let P, Q, R be probability distributions. Suppose* $\mathrm{D}_\alpha(P\|R) \le a\cdot\alpha$ *and* $\mathrm{D}_\alpha(R\|Q) \le b\cdot\alpha$ *for all* $\alpha \in (1,\infty)$*. Then, for all* $\alpha \in (1,\infty)$,

$$\mathrm{D}_\alpha(P\|Q) \le \alpha\cdot(\sqrt{a}+\sqrt{b})^2 \le 2\alpha\cdot(a+b).$$

- Decompose what we want to bound

$$\mathrm{D}_\alpha\left(Z\|e^t Z + s\right)$$

$$\mathrm{D}_\alpha\left(e^t Z + s\|Z\right)$$

Bounding the two parts separately

Lemma 19. *Let* $Z \leftarrow \mathsf{LLN}(\sigma)$ *for* $\sigma > 0$. *Let* $t \in \mathbb{R}$ *and* $\alpha \in (1, \infty)$. *Then*

$$\mathrm{D}_\alpha\left(Z \middle\| e^t Z\right) \le \frac{\alpha t^2}{2\sigma^2}.$$

- Proof:

$$\mathrm{D}_\alpha\left(Z \middle\| e^t Z\right) = \mathrm{D}_\alpha\left(X e^{\sigma Y} \middle\| X e^{\sigma Y + t}\right) \le \sup_x \mathrm{D}_\alpha\left(x e^{\sigma Y} \middle\| x e^{\sigma Y + t}\right) \le \mathrm{D}_\alpha\left(\sigma Y \| \sigma Y + t\right).$$

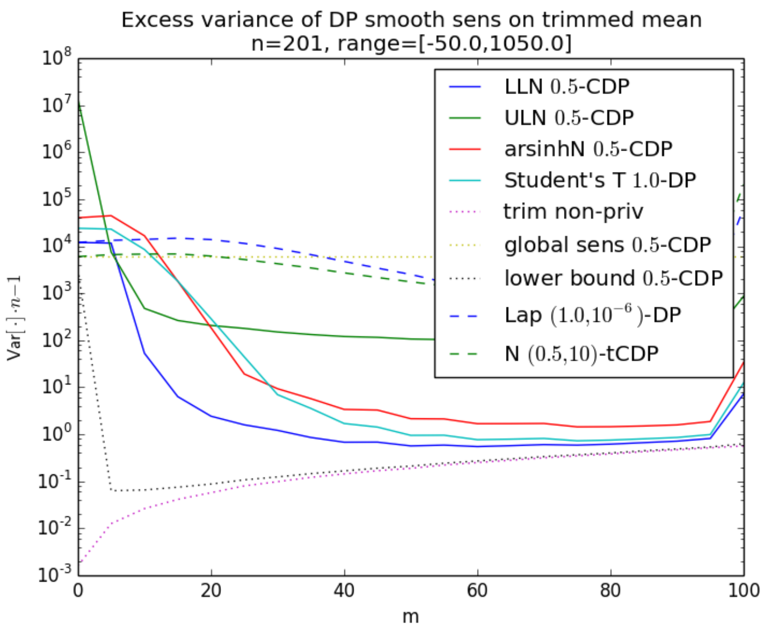
Lemma 20. *Let* $Z \leftarrow \mathsf{LLN}(\sigma)$ *for* $\sigma > 0$. *Let* $s \in \mathbb{R}$ *and* $\alpha \in (1, \infty)$. *Then*

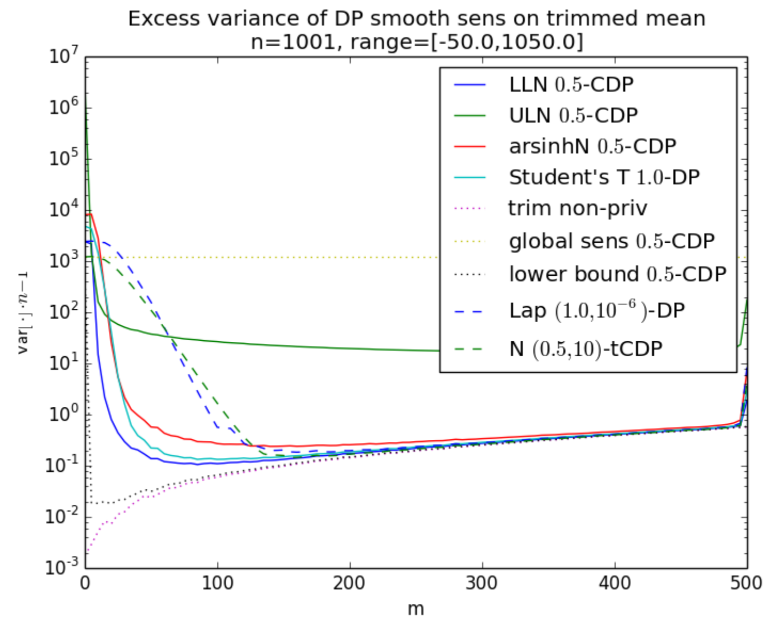
$$\mathrm{D}_\alpha\left(Z \| Z + s\right) \le \min\left\{\frac{1}{2} e^{3\sigma^2} s^2 \alpha, e^{\frac{3}{2}\sigma^2} s\right\}.$$

- Proof:

Improvement from running smoothed sensitivity is substantial!

$$\text{trim}_m(x) = \frac{x_{(m+1)} + x_{(m+2)} + \cdots + x_{(n-m)}}{n-2m},$$





Bun and Steinke (2019): "Average case averages": https://arxiv.org/pdf/1906.02830.pdf

Next lecture

- Propose-Test-Release
- Stability-based query release
- Application to PATE