CSC2541 Lecture 6
– Learning with differential privacy
– Statistical query and local model equivalence
– Factorization mechanism

Alex Edmonds

October 15, 2019

University of Toronto

Differential privacy

Definition

Datasets $X, Y \in \mathbb{N}^{\mathcal{X}}$ are called adjacent if

$$||X - Y||_1 \leq 1$$

(different by the addition or removal of a single datapoint).

Definition

A randomized function $\mathcal{M} : \mathbb{N}^{\mathcal{X}} \to \Omega$ is ε-differentially private if, for all adjacent $X, Y \in \mathbb{N}^{\mathcal{X}}$, every outcome $S \subseteq \Omega$ satisfies,

$$\Pr_{\mathcal{M}}[\mathcal{M}(X) \in S] \leq e^{\varepsilon} \cdot \Pr_{\mathcal{M}}[\mathcal{M}(Y) \in S].$$

"No event is made much more (or less) likely by my participation."

Exponential mechanism (private optimization)

Given a dataset $X \in \mathbb{N}^{\mathcal{X}}$, utility of outcome $r \in S$ is $u(X, r) \in \mathbb{R}$.

Sensitivity is given by

$$\Delta u = \max_{r \in S} \max_{X,Y:||X-Y||_1 \leq 1} |u(X,r) - u(Y,r)|$$

Exponential mechanism optimizes $u(X, r)$ by selecting $r^* := \mathcal{M}(X)$ with

$$\Pr_{\mathcal{M}}[\mathcal{M}(X) = r] \propto \exp\left(\frac{\varepsilon u(X,r)}{2\Delta}\right)$$

- satisfies ε-DP
- accuracy guaranteed according to

$$\Pr_{r^* \sim \mathcal{M}(X)}\left[u(X, r^*) \leq OPT_u(X) - \frac{2\Delta}{\varepsilon}\left(\log\left(\frac{|S|}{|S_{OPT}(X)|}\right)\right)\right]$$

where $OPT_u(X) = \max_r u(X, r)$,
$S_{OPT}(X) = \{r \in S \ : \ u(X, r) = OPT_u(X)\}$.

PAC learning (probably approximately correct)

Definition

Say $\mathcal{M}$ that $\underline{(\alpha, \beta)\text{-PAC learns}}$ concept class $\mathcal{C}$ if,

for every distribution $\mathcal{P}$ on inputs,
for every labelling function $f \in \mathcal{C}$, $f : \mathcal{X} \to \{-1, +1\}$,
given samples $x_1, x_2, \ldots, x_n \overset{iid}{\sim} \mathcal{P}$ labelled as

$$(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_m, f(x_m)),$$

$\mathcal{M}$ produces f^* such that

$$Pr_{x_1,\ldots,x_n,\mathcal{M}}[A(f^*, \mathcal{P}) \geq 1 - \alpha] \geq 1 - \beta$$

where $A(f^*, \mathcal{P}) = \Pr_{x \sim \mathcal{P}}[f^*(x) = f(x)]$.

Given $x_1, \ldots, x_n$ and $f(x_1), \ldots, f(x_n)$,
natural approach is to take $f^* = \underbrace{\arg\max_{f \in \mathcal{C}} \sum_i \mathbb{I}[f^*(x_i) = f(x_i)]}_{\text{empirical accuracy maximizer}}$.

Private PAC learning – sample complexity

Without privacy:

"Given $x_1, \ldots, x_n$ and $f(x_1), \ldots, f(x_n)$, take $f^* = \arg\max_f \sum_i \mathbb{I}[f^*(x_i) = f(x_i)]$."

i.e., maximize $u(X, f) := \sum_i \mathbb{I}[f^*(x_i) = f(x_i)]$. $[X = \{(x_1, f(x_1)), \ldots, (x_n, f(x_n))\}]$

To $\underline{(\alpha, 1/4)\text{-PAC learn } \mathcal{C}}$ requires at most

$$n = O\left(\frac{\log |\mathcal{C}|}{\alpha^2}\right) \text{ samples.}$$

With privacy: [KLN+11]

Apply exponential mechanism to u to learn privately!

To $\underline{(\alpha, 1/4)\text{-PAC learn } \mathcal{C} \text{ with } \varepsilon\text{-differential privacy}}$ requires at most

$$n = O\left(\max\left\{\frac{\log |\mathcal{C}|}{\varepsilon\alpha}, \frac{\log |\mathcal{C}|}{\alpha^2}\right\}\right) \text{ samples.}$$

Good news! ε-DP for free when $\varepsilon \geq \alpha$.

Private PAC learning – sample complexity

Without privacy:

VC-dimension bounds

To $(\alpha, 1/4)$-PAC learn $\mathcal{C}$ requires at most

$$n = O\left(\frac{VC(\mathcal{C})}{\alpha^2}\right) = O\left(\frac{\log|\mathcal{C}|}{\alpha^2}\right) \text{ samples}$$

where $VC(\mathcal{C})$ is the VC-dimension of $\mathcal{C}$.

There exist $\mathcal{C}$ where $VC(C) \ll \log|\mathcal{C}|$.
$\Rightarrow$ Much better learning guarantees.

With privacy: [KLN+11]

To $(\alpha, 1/4)$-PAC learn $\mathcal{C}$ with ε-differential privacy requires at most

$$n = O\left(\max\left\{\frac{VC(\mathcal{C}) \cdot \log|\mathcal{X}|}{\varepsilon\alpha}, \frac{VC(\mathcal{C}) \cdot \log|\mathcal{X}|}{\alpha^2}\right\}\right) \text{ samples.}$$

Dependence on $\log|\mathcal{X}|$ is necessary for private learners.

Private PAC learning – computational complexity

$\boxed{\text{Caveats } ☹}$

Recall, the exponential mechanism selects r^* according to

$$\Pr_{\mathcal{M}}[\mathcal{M}(X) = f] \propto \exp\left(\frac{\varepsilon u(X, r)}{2\Delta}\right)$$

$$\Rightarrow$$

$$\Pr_{\mathcal{M}}[\mathcal{M}(X) = f] = \frac{\exp\left(\frac{\varepsilon u(X,f)}{2\Delta}\right)}{\sum_{f \in \mathcal{C}} \exp\left(\frac{\varepsilon u(X,f)}{2\Delta}\right)}$$

However, computing the denominator can be computationally expensive if $|\mathcal{C}|$ is large.

For learning,

- computational complexity of exp. mechanism: $\geq$ linear in $|\mathcal{C}|$
- sample complexity of exp. mechanism: logarithmic in $|\mathcal{C}|$

Typically, $|\mathcal{C}| \geq 2^{\Omega(d)}$ where d is dimension.

Computationally efficient learning with statistical queries

Statistical queries [DR14]

The statistical query model of machine learning restricts the learner's distribution access to an *oracle* which,

$$\text{given} \quad \phi : \mathcal{X} \times \{-1, 1\} \to [-1, 1],$$

$$\text{returns } z, \quad \text{where} \quad |\mathbb{E}_{x \sim \mathcal{P}} \phi(x, f(x)) - z| \leq \alpha.$$

Since each ϕ has *sensitivity*

$$\Delta\phi := \max_{||X-Y||_1=1} ||\phi(x) - \phi(x)|| = 1/n,$$

they can be answered with the *Laplace mechanism.* (previous lecture)

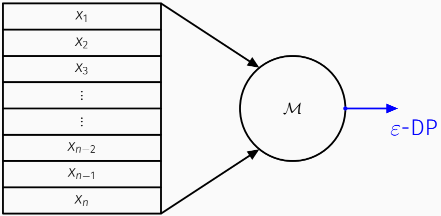
$\Rightarrow$ Learning algorithm which requires answers to m SQs can be simulated with ε-DP given

$$n = O\left(\max\left\{\frac{m \log m}{\varepsilon\alpha}, \frac{m \log m}{\alpha^2}\right\}\right) \text{ samples.}$$

Useful if $m \log m < |\mathcal{C}|$. ☺

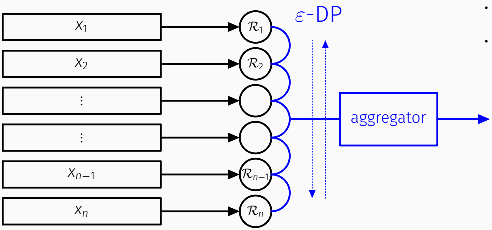
Central model / local model

Central model (DP)



- trusted central coordinator
- only final output is ε-DP

Local model (LDP)



- each row held by a distinct agent
- entire transcript must be ε-DP

Local differential privacy

We have seen LDP before:

Randomized response

If each agent i reports their bit $b_i \in \{0, 1\}$ as $\tilde{b}_i$

- truthfully, with probability $\frac{1+\varepsilon}{2}$,
- falsely, with probability $\frac{1-\varepsilon}{2}$.

Even if every bit $\tilde{b}_i$ is released, ε-DP is preserved.

Local differential privacy

General local protocols

A local protocol may be obtained
by having agents communicate in rounds.

Each round t is assigned a privacy parameter ε_t.

In round t, each agent reports on their sample x_i with ε_t-DP.
$\Rightarrow$ Entire transcript is ε-DP for $\varepsilon := \sum_t \varepsilon_t$.

Not the only way of obtaining local privacy [JMNR19],
but we restrict ourselves to protocols such as these.

SQ and local models

For local DP,
the Laplace mechanism can be applied to a single sample.

Local Laplace mechanism

Consider a dataset $X = \{x\}$ of one sample.

The ℓ_1 sensitivity of $\phi : \mathcal{X} \to [-1, +1]$ is

$$\Delta\phi = \max_{x,y\in\mathcal{X}} ||\phi(x) - \phi(y)||_1 = 1$$

so ε-DP is satisfied when an agent releases

$$\mathcal{R}(x) = \phi(x) + w \text{ where } w \sim \mathrm{Lap}(\Delta\phi/\varepsilon).$$

When $x_1, \ldots, x_n \overset{iid}{\sim} \mathcal{P}$, then $z := \frac{1}{n}\sum_{i\in[n]} \mathcal{R}(x_i)$ satisfies

$$\Pr[|z - \mathbb{E}_{b\sim\mathcal{P}}[\phi(b)]| \le \tau] \ge \frac{3}{4} \quad \text{with } n = O\left(\frac{1}{\tau^2\varepsilon^2}\right) \text{ samples.}^*$$

*Note: $O(\frac{1}{\tau\varepsilon})$ samples suffice for Laplace in the central model.

SQ-algorithm $\Rightarrow$ Local protocol

Answering SQs with the local Laplace mechanism
allows simulation of any SQ-learner in the local model:

Theorem [KLN+11]

Any SQ-algorithm $\mathcal{A}$ which
(α, β)-PAC learns $\mathcal{C}$ with m SQs of accuracy τ,
may be used to obtain
a locally ε-DP protocol which $(\alpha, 2\beta)$-PAC learns $\mathcal{C}$ with

$$n = O\left(\frac{m \log(m/\beta)}{\varepsilon^2 \tau^2}\right) \quad \text{samples.}$$

A convenient way to obtain
locally private protocols!

Local protocol ⇒ SQ-algorithm

Theorem [KLN+11]

Any locally ε-DP protocol $\mathcal{M}$ on n agents which (α, β)-PAC learns $\mathcal{C}$ may be used to obtain
an SQ-algorithm $\mathcal{A}$ which $(\alpha, 2\beta)$-PAC learns $\mathcal{C}$
by making $O(n \cdot e^{\varepsilon})$ SQs of accuracy $\tau = \Theta(\beta/(e^{2\varepsilon} n))$.

Proof idea.

Given ε-DP $\mathcal{R} : \mathcal{X} \to \mathcal{Z}$, want to sample $z \in \mathcal{Z}$ with prob. $p(z) = \Pr_{\mathcal{R},x}[\mathcal{R}(x) = z]$.

Uses fact that $p(z)$ can be approximated with SQ since

$$p(z) = \mathbb{E}_x[\phi(x)] \quad \text{where} \quad \phi(x) := \Pr_{\mathcal{R}}[\mathcal{R}(x) = z].$$

Apply approximate version of following rejection sampling strategy:

1. Sample z with probability $\Pr_{\mathcal{R}}[\mathcal{R}(0) = z]$;
2. Accept z with probability $\Pr_{\mathcal{R},x}[\mathcal{R}(x) = z]/(e^{\varepsilon} \cdot \Pr_{\mathcal{R}}[\mathcal{R}(0) = z])$.
   - gives much better guarantee than using approximation of $p(z)$ directly
   - $\Pr_{\mathcal{R},x}[\mathcal{R}(x) = z]/\Pr_{\mathcal{R}}[\mathcal{R}(0) = z] \leq e^{\varepsilon} \Rightarrow O(e^{\varepsilon})$ rejections expected

□

SQ and local models

Local protocol $\Rightarrow$ SQ-algorithm

SQ lower bound $\Rightarrow$ Local model lower bound

Definition

A parity $f_S : \{-1, 1\}^d \to \{-1, 1\}$ is determined by a set $S \subset [d]$ with

$$f_S(x) := \prod_{i \in S} x_i$$

Let PARITY_d be the class of all such functions.

Theorem [BFJ+94]

No SQ-learner for PARITY_d using at most $2^{d/3}$ SQs of accuracy $2^{-d/3}$.

Corollary [KLN+11]

For constant ε, for all d, for some $n = 2^{\Omega(d)}$,
there exists no locally ε-DP learner for PARITY_d on n agents.

SQ and local models

Corollary [KLN+11]

For constant ε, for all d, for some $n = 2^{\Omega(d)}$,
there exists no locally ε-DP learner for PARITY_d on n agents.

vs.

Theorem [KLN+11]

In the central model, there exists an ε-DP mechanism $\mathcal{M}$ which (α, β)-PAC learns PARITY_d with $n = O(\frac{d \log(1/\beta)}{\varepsilon\alpha})$ samples.

($\mathcal{M}$ can even be made computationally efficient)

Exponential separation between
local and central models!

Factorization mechanism

Estimating SQs with LDP is building block of LDP mechanisms.
$\Rightarrow$ Want to answer SQs in local model with optimal sample-efficiency.

(Let $\mathcal{X} = [N]$.)

Given $\phi_1, \ldots, \phi_m$, where $\phi_j : \mathcal{X} \to [-1, 1]$, want estimates $z_1, \cdots, z_m$ s.t.

$$|\mathbb{E}_{x\sim\mathcal{P}}[\phi_j(x)] - z_j| \leq \tau \qquad \forall j$$

when each agent i gets $x_i \sim \mathcal{P}$, and $\mathcal{P}$ is unknown distribution on $\mathcal{X}$.

Equivalently,

- consider matrix $W \in \mathbb{R}^{m\times N}$ given by $w_{j,k} = \phi_j(k)$;
- let $h_{\mathcal{P}} := (\mathcal{P}(1), \ldots, \mathcal{P}(m))$;

and estimate $Wh_{\mathcal{P}} \in \mathbb{R}^m$ by $\mathcal{M}(X) \in \mathbb{R}^m$ so that

$$|Wh_{\mathcal{P}} - \mathcal{M}(X)|_\infty \leq \tau.$$

Factorization mechanism

We saw that $n = O\left(\frac{m \log m}{\varepsilon^2 \tau^2}\right)$ samples sufficed to obtain, w.h.p.,

$$|Wh_{\mathcal{P}} - \mathcal{M}(X)|_\infty \leq \tau.$$

However, naivly estimating $\mathbb{E}_{x \sim \mathcal{P}}[\phi_i(x)]$ separately for each ϕ_i can be badly sub-optimal.

Example: Repeated queries

If $\phi_1 = \cdots = \phi_m$, then

$$n = O\left(\frac{1}{\varepsilon^2 \tau^2}\right)$$

samples suffice to obtain $z = \mathbb{E}_{x \sim \mathcal{P}}[\phi_1(x)] \pm \tau$.

Then z can be reused to answer each of the queries $\phi_2 \cdots \phi_m$.

Factorization mechanism

Example: Threshold queries (estimating the CDF of $\mathcal{P}$)

The set of threshold queries $\{\phi_j\}_{j\in[N]}$ on $[N]$ is given by

$$\phi_j(x) = \begin{cases} 1 & x \le j \\ 0 & \text{otherwise} \end{cases}$$

Strategy: Factor W as $W = RA$ where R and A are also matrices.

Then $Wh_{\mathcal{P}} = R(Ah_p)$.

So, obtain an estimate Z of Ah_p in the local model. Then return RZ.

One such factorization gives A which corresponds to queries

$$\left\{ \begin{array}{l} \phi_{1:N}(x) \\ \phi_{1:N/2}(x), \phi_{N/2+1:N}(x) \\ \phi_{1:N/4}(x), \phi_{1:N/4}(x), \cdots, \phi_{3N/4+1:N}(x) \\ \vdots \\ \phi_{0:1}(x), \phi_{1:2}(x), \cdots, \phi_{N-1:N}(x) \end{array} \right\} \quad \text{where} \quad \phi_{s:t}(x) := \begin{cases} 1 & s < x \le j \\ 0 & \text{otherwise} \end{cases}$$

Answers to these queries allow us to reconstruct an answer for each ϕ_i, i.e. $\mathbb{E}_{X\sim\mathcal{P}}\phi_7(X) = \mathbb{E}_{X\sim\mathcal{P}}\phi_{1:4}(X) + \mathbb{E}_{X\sim\mathcal{P}}\phi_{5:6}(X) + \mathbb{E}_{X\sim\mathcal{P}}\phi_{6:7}(X)$

Factorization mechanism

The factorization strategy may be generally applied.

For a factorization $W = RA$, we may bound the error of our mechanism by

$$\frac{||R||_{2\to\infty}||A||_{1\to 2}\sqrt{\log m}}{\varepsilon\sqrt{n}}$$

where $||\cdot||_{2\to\infty}$ and $||\cdot||_{1\to 2}$ are matrix operator norms.

This motivates us to minimize $||R||_{2\to\infty}||A||_{1\to 2}$ subject to $W = RA$.

Let $\gamma_2(W) := \min\{||R||_{2\to\infty}||A||_{1\to 2}\}$.

In particular, error τ may be obtained with $n = O\left(\frac{\gamma_2(W)^2 \log m}{\varepsilon^2\tau^2}\right)$ samples.

Avrim Blum, Merrick L. Furst, Jeffrey C. Jackson, Michael J. Kearns, Yishay Mansour, and Steven Rudich.
Weakly learning DNF and characterizing statistical query learning using fourier analysis.
In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, 23-25 May 1994, Montréal, Québec, Canada*, pages 253–262, 1994.

Cynthia Dwork and Aaron Roth.
The algorithmic foundations of differential privacy.
Foundations and Trends in Theoretical Computer Science, 9(3-4):211–407, 2014.

Matthew Joseph, Jieming Mao, Seth Neel, and Aaron Roth.
The role of interactivity in local differential privacy.
CoRR, abs/1904.03564, 2019.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith.
What can we learn privately?
SIAM J. Comput., 40(3):793–826, 2011.