DIFFERENTIALLY PRIVATE MACHINE LEARNING

Aurélien Bellet (Inria)

Journées de Statistique (JDS)
June 13, 2022
PRIVACY IN THE BIG DATA ERA

• Massive collection of personal data by companies and public organizations, driven by the progress of data science and AI

• Data is increasingly sensitive and detailed: browsing history, purchase history, social network posts, speech, geolocation, health...

• Quantifying privacy risks is challenging!
  • Attacker may have prior knowledge
  • Same data used in multiple computations
  • Indirect leakage from aggregate quantities
• Aggregate (potentially noisy) statistics about many individuals are vulnerable to various attacks on data privacy

• **Membership inference attacks**, i.e. inferring presence of known individual in a dataset from (high-dimensional) aggregate statistics
  • Example: statistics about genomic variants [Homer et al., 2008]

• **Reconstruction attacks**, i.e. inferring (part of) the dataset from the output of many aggregate statistics
  • After sufficiently many queries, one can reconstruct the dataset [Dinur and Nissim, 2003]
MACHINE LEARNING MODELS ARE NOT SAFE

• Machine learning models are elaborate kinds of aggregate statistics

• They are also susceptible to membership inference and reconstruction attacks, see e.g. [Shokri et al., 2017, Paige et al., 2020, Geiping et al., 2020]
• **Goal:** achieve utility while preserving privacy (conflicting objectives!)

• Note: this is separate from security concerns (e.g., unauthorized access to the system)
1. Differential Privacy

2. Private learning in the centralized setting

3. Private learning without a trusted curator
Differential Privacy
Differential Privacy

Randomized algorithm $A(x_1, x_2, \ldots, x_n)$ with random coins

$A(D)$ distribution of $A(D)$

$A(D')$ distribution of $A(D')$

- Neighboring datasets $D = \{x_1, x_2, \ldots, x_n\}$ and $D' = \{x_1', x_2, x_3, \ldots, x_n\}$
- Requirement: $A(D)$ and $A(D')$ should have "close" distribution

Probability ratio bounded

Output range of $A$
**Definition ([Dwork et al., 2006], informal)**

A randomized algorithm $\mathcal{A}$ is $(\varepsilon, \delta)$-differentially private (DP) if for all neighboring datasets $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ and $\mathcal{D}' = \{x_1, x'_2, x_3, \ldots, x_n\}$ and all sets $S$:

$$\Pr[\mathcal{A}(\mathcal{D}) \in S] \leq e^\varepsilon \Pr[\mathcal{A}(\mathcal{D}') \in S] + \delta.$$ 

- DP is a property of the analysis, not of a particular output.
- Sufficient condition: for $o \sim \mathcal{A}(\mathcal{D})$, the privacy loss $\ln \left( \frac{\Pr[\mathcal{A}(\mathcal{D}) = o]}{\Pr[\mathcal{A}(\mathcal{D}') = o]} \right)$ is bounded by $\varepsilon$ with probability $1 - \delta$ (note: $\varepsilon$ can be seen as a function of $\delta$).
- For meaningful privacy guarantees, think of $\varepsilon \leq 1$ and $\delta \ll 1/n$.
- In 2017, Dwork, McSherry, Nissim & Smith won the Gödel prize for introducing DP.
- In 2020, the US Census started to use DP for its data releases.
PROPERTIES OF DP: ROBUSTNESS TO POSTPROCESSING AND AUXILIARY KNOWLEDGE

- **Robustness to processing**: informally, if \( A \) is \((\epsilon, \delta)\)-DP, then so is \( f \circ A \) for any \( f \)

- **Robustness to auxiliary knowledge**: DP bounds the relative advantage that an adversary gets from observing the output of an algorithm
  - DP holds even if adversary knows all but one data record
  - Interpretation as hypothesis testing: adversary knows \( A \) and neighboring datasets \( D_0 \) and \( D_1 \), observes a realization of \( A(D_b) \) for a secret bit \( b \in \{0, 1\} \), and must guess whether it was drawn from \( A(D_0) \) or \( A(D_1) \)
  - DP puts a bound on the trade-offs between the true positive rate and the false positive rate that can be achieved for this test
PROPERTIES OF DP: COMPOSITION

• Composition allows to control the worst-case cumulative privacy loss over multiple analyses run on the same dataset, including complex multi-step algorithms

Theorem (Simple composition)

Let $A_1, \ldots, A_K$ be such that $A_k$ satisfies $(\epsilon_k, \delta_k)$-DP. For any dataset $D$, let $A$ be such that $A(D) = (A_1(D), \ldots, A_k(D))$. Then $A$ is $(\epsilon, \delta)$-DP with $\epsilon = \sum_{k=1}^{K} \epsilon_k$ and $\delta = \sum_{k=1}^{K} \delta_k$.

Theorem (Advanced composition)

Let $\epsilon, \delta, \delta' > 0$. If $A_k$ satisfies $(\epsilon, \delta)$-DP, then $A$ is $(\epsilon', K\delta + \delta')$-DP with

$$\epsilon' = \sqrt{2K \ln(1/\delta')} \epsilon + K\epsilon(e^\epsilon - 1)$$

• The sequence of algorithms can be chosen adaptively

• Numerically tighter composition can be obtained with through a variant of DP based on the Rényi divergence [Mironov, 2017]
• Consider \( f \) taking as input a dataset and returning a \( p \)-dimensional real vector

Gaussian mechanism \( \mathcal{A}_{\text{Gauss}}(\mathcal{D}, f, \varepsilon, \delta) \)

1. Compute sensitivity \( \Delta = \max_{\mathcal{D}, \mathcal{D}'} \text{ are neighboring } \| f(\mathcal{D}) - f(\mathcal{D'}) \|_2 \)

2. Output \( f(\mathcal{D}) + \eta \), where \( \eta \sim \mathcal{N}(0, \sigma^2 I_p) \) with \( \sigma = \frac{\sqrt{2 \ln(1.25/\delta)} \Delta}{\varepsilon} \)

Theorem

Let \( \varepsilon, \delta > 0 \). The Gaussian mechanism \( \mathcal{A}_{\text{Gauss}}(\cdot, f, \varepsilon, \delta) \) is \( (\varepsilon, \delta) \)-DP.

• Noise calibrated using sensitivity of \( f \) and privacy budget \( (\varepsilon \text{ and } \delta) \)

• Sketch of proof: tail bound for the Gaussian distribution + simplifications

• DP induces a privacy-utility trade-off, here in terms of the variance of the estimate

• Note: the MSE achieved by the Gaussian mechanism is worst-case optimal
PRIVATE LEARNING IN THE CENTRALIZED SETTING
A trusted curator wants to privately release a model trained on data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$.

We focus here on approximately solving an Empirical Risk Minimization (ERM) problem under an $(\epsilon, \delta)$-DP constraint:

$$\min_{\theta \in \Theta} \left\{ F(\theta; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} L(\theta; x_i, y_i) \right\}$$

(Note: in some cases, DP can imply generalization [Bassily et al., 2016, Jung et al., 2021])

We can achieve this by designing a differentially private ERM solver.
Algorithm: Differentially Private SGD $A_{\text{DP-SGD}}(\mathcal{D}, L, \epsilon, \delta)$

- Initialize parameters to $\theta^{(0)} \in \Theta$ (must be independent of $\mathcal{D}$)
- For $t = 0, \ldots, T - 1$:
  - Pick random mini-batch $B^{(t)} \subseteq \{1, \ldots, n\}$ of size $m$
  - $\eta^{(t)} \leftarrow (\eta_1^{(t)}, \ldots, \eta_p^{(t)}) \in \mathbb{R}^p$ where each $\eta_j^{(t)} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = \frac{16l\sqrt{T\ln(2/\delta)\ln(1.25T/\delta n)}}{n\epsilon}$
  - $\theta^{(t+1)} \leftarrow \Pi_\Theta \left( \theta^{(t)} - \gamma_t (\nabla L(\theta^{(t)}; B^{(t)}) + \eta^{(t)}) \right)$ (\(\Pi_\Theta\) projection operator)
- Return $\theta^{(T)}$

- More data (larger $n$) → less noise added to each gradient
- More iterations (larger $T$) → more noise added to each gradient

Theorem (DP guarantees for DP-SGD)

Let $\epsilon \leq 1, \delta > 0$. Let the loss function $L(\cdot; x, y)$ be $l$-Lipschitz w.r.t. the $\ell_2$ norm for all $x, y \in \mathcal{X} \times \mathcal{Y}$. Then $A_{\text{DP-SGD}}(\cdot, L, \epsilon, \delta)$ is $(\epsilon, \delta)$-DP.
Sketch of proof.

• Recall that for a query with $\ell_2$ sensitivity $\Delta$, achieving $(\epsilon', \delta')$ with the Gaussian mechanism requires to add noise with standard deviation $\sigma' = \sqrt{\frac{2 \ln(1.25/\delta')}{\epsilon'}} \Delta$

• The loss function $L$ is $l$-Lipschitz, which implies that $\ell_2$-norm of individual gradients is bounded by $l$ and therefore $\Delta = 2l/m$

• Hence, with $\sigma = \frac{16l\sqrt{T\ln(2/\delta)\ln(1.25T/\delta n)}}{n\epsilon}$, each noisy gradient is $\left(\frac{n\epsilon}{4m\sqrt{2T\ln(2/\delta)}}, \frac{\delta n}{2mT}\right)$-DP

• Using privacy amplification by subsampling [Balle et al., 2018] allows to leverage the randomness in the choice of $\mathcal{B}$: each noisy gradient is in fact $\left(\frac{\epsilon}{2\sqrt{2T\ln(2/\delta)}}, \frac{\delta}{2T}\right)$-DP

• DP-SGD is an adaptive composition of $T$ DP mechanisms, so by advanced composition we obtain that it is $(\epsilon, \delta)$-DP
Theorem (Utility guarantees for DP-SGD [Bassily et al., 2014])

Let $\Theta$ be a convex domain of diameter bounded by $R$, and let the loss function $L$ be convex and $l$-Lipschitz over $\Theta$. For $T = n^2$ and $\gamma_t = O(R/\sqrt{t})$, DP-SGD guarantees:

$$\mathbb{E}[F(\theta^{(T)})] - \min_{\theta \in \Theta} F(\theta) \leq O\left(\frac{lR \sqrt{p \ln(1/\delta)} \ln^{3/2}(n/\delta)}{n\epsilon}\right).$$

- Proof: plug variance of stochastic gradients in analysis of SGD [Shamir and Zhang, 2013]
- Utility gap w.r.t. the non-private model is $\tilde{O}(\sqrt{p}/\epsilon n)$, which is worst-case optimal
- In practice: drop Lipschitz assumption and use gradient clipping [Abadi et al., 2016], which introduces a bias-variance trade-off in gradient estimation
PRIVATE LEARNING WITHOUT A TRUSTED CURATOR
• In the real world data is often decentralized across different parties.

• Data may be considered too sensitive to be shared (e.g., due to legal restrictions, intellectual property rights, or because it provides a competitive advantage).

• Inferior performance and/or biased results if each party learns independently.
Federated Learning (FL) aims to collaboratively train ML models while keeping the data decentralized.

- FL is a **booming** and **multidisciplinary** topic: see collaborative survey [Kairouz et al., 2021] to know more about existing work and open problems.

- FL does not itself provide any privacy guarantees: in fact, it offers an additional attack surface compared to the centralized setting as participants will observe some intermediate results [Nasr et al., 2019, Geiping et al., 2020].
Central DP: a trusted curator collects raw data and runs a DP algorithm on it

→ the observed output is only the final result

Local DP: no trusted curator so each party must locally run a DP algorithm

→ the observed output consists of all messages shared by all parties
A KEY FUNCTIONALITY FOR FL: DP AGGREGATION

• Consider $K$ parties, with each party $k$ holding local dataset $\mathcal{D}_k$

• Many FL algorithms rely on a coordinating server and proceed as follows:

  \[
  \text{for } t = 1 \text{ to } T \text{ do} \\
  \hspace{1cm} \text{At each party } k: \text{ compute } \theta_k \leftarrow \text{LOCALUPDATE}(\theta, \theta_k; \mathcal{D}_k), \text{ send } \theta_k \text{ to server} \\
  \hspace{1cm} \text{At server: compute } \theta \leftarrow \frac{1}{K} \sum_k \theta_k, \text{ send } \theta \text{ back to the parties}
  \]

• Therefore: DP aggregation + Composition property of DP $\implies$ DP-FL

• **DP aggregation**: given a private value $\theta_k \in [0, 1]$ for each party $k$, we want to accurately estimate $\theta_{\text{avg}} = \frac{1}{K} \sum_k \theta_k$ under a DP constraint

• **Central DP**: trusted server computes $\theta_{\text{avg}}$ and adds Gaussian noise

• **Local DP**: each party $k$ adds Gaussian noise to $\theta_k$ before sharing it

\[\text{Error is } \sqrt{K} \text{ larger in local DP } \implies \text{ study intermediate trust models}\]
Assume that pairs of parties can communicate through secure channels (the server may serve as relay), e.g. using a public key infrastructure

**Algorithm** GOPA protocol [Sabater et al., 2020]

- Each party $k$ generates independent Gaussian noise $\eta_k$
- Each party $k$ selects a random set of $m$ other parties
- for all selected pairs of parties $k \sim l$ do
  - Parties $k$ and $l$ securely exchange pairwise-canceling Gaussian noise $\Delta_{k,l} = -\Delta_{l,k}$
  - Each party $k$ sends $\hat{\theta}_k = \theta_k + \sum_{k \sim l} \Delta_{k,l} + \eta_k$ to the server

- **Estimate of the average:** $\hat{\theta}^{avg} = \frac{1}{K} \sum_k \hat{\theta}_k = \theta^{avg} + \frac{1}{K} \sum_k \eta_k$

- Intuition: pairwise noise does not affect utility but helps protecting individual values
Privacy Guarantees for GOPA

- **Adversary**: coalition of the server with a proportion $1 - \tau$ of the parties

**Theorem (Privacy of GOPA [Sabater et al., 2020], informal)**

- Let each party select $m = O(\log(\tau K)/\tau)$ other parties
- Set the independent noise variance so as to satisfy $(\epsilon, \delta')$-DP in the central model
- For large enough pairwise noise variance, GOPA is $(\epsilon, \delta)$-DP with $\delta = O(\delta')$.

- Same utility as central DP with only logarithmic number of messages per party
- Our theoretical results give practical values for the quantities above
- More generally, we precisely quantify the effect of the graph of communications between honest parties on the privacy guarantees
For reasonable proportions $\rho$ of honest parties, the variance of the estimated average produced by GOPA is similar to the trusted curator setting.

As expected, the resulting FL model also has similar accuracy.
• In fully decentralized FL, global aggregations are replaced by local aggregations among neighbors in a graph (thus, the previous approach cannot be applied).

• But there is no server observing all messages, and each party $k$ has a limited view.

• Can this be used to prove stronger differential privacy guarantees?
• Let $O_k$ be the set of messages sent and received by party $k$

**Definition (Network DP [Cyffers and Bellet, 2022])**

An algorithm $A$ satisfies $(\epsilon, \delta)$-network DP if for all pairs of distinct parties $k, l \in \{1, \ldots, K\}$ and all pairs of datasets $D, D'$ that differ only in the local dataset of party $l$, we have:

$$\Pr[O_k(A(D))] \leq e^\epsilon \Pr[O_k(A(D'))] + \delta.$$

• This is a relaxation of local DP: if $O_k$ contains the full transcript of messages, then network DP boils down to local DP
• Consider the standard objective $F(\theta; D) = \frac{1}{K} \sum_{k=1}^{K} F_k(\theta; D_k)$ and a complete graph

• We consider a fully decentralized algorithm where the model is updated sequentially by following a random walk

**Algorithm**  
Private decentralized SGD on a complete graph

**Initialize model** $\theta$

**for** $t = 1$ to $T$ **do**

  Current party updates $\theta$ by a gradient update with Gaussian noise

  Current party sends $\theta$ to a random party

**return** $\theta$
Consider the standard objective $F(\theta; \mathcal{D}) = \frac{1}{K} \sum_{k=1}^{K} F_k(\theta; \mathcal{D}_k)$ and a complete graph.

We consider a fully decentralized algorithm where the model is updated sequentially by following a random walk.

**Algorithm** Private decentralized SGD on a complete graph

1. Initialize model $\theta$
2. For $t = 1$ to $T$
   - Current party updates $\theta$ by a gradient update with Gaussian noise
   - Current party sends $\theta$ to a random party
3. Return $\theta$
WALK-BASED DECENTRALIZED SGD

- Consider the standard objective $F(\theta; D) = \frac{1}{K} \sum_{k=1}^{K} F_k(\theta; D_k)$ and a complete graph.
- We consider a fully decentralized algorithm where the model is updated sequentially by following a random walk.

**Algorithm** Private decentralized SGD on a complete graph

1. Initialize model $\theta$.
2. **for** $t = 1$ to $T$ **do**
   - Current party updates $\theta$ by a gradient update with Gaussian noise.
   - Current party sends $\theta$ to a random party.
3. **return** $\theta$.
Theorem ([Cyffers and Bellet, 2022], informal)

To achieve a fixed $(\epsilon, \delta)$-DP guarantee with the previous algorithm, the standard deviation of the noise is $O(\sqrt{K}/\ln K)$ smaller under network DP than under local DP.

- Accounting for the limited view in fully decentralized algorithms amplifies privacy guarantees by a factor of $O(\ln K/\sqrt{K})$, nearly recovering the utility of central DP.

- The proof leverages recent results on privacy amplification by iteration [Feldman et al., 2018] and exploits the randomness of the path taken by the model.

- We show some robustness to collusion (albeit with smaller privacy amplification).
• Results are consistent with our theory: network DP-SGD significantly amplifies privacy guarantees compared to local DP-SGD
Wrapping up
• **Differential privacy** provides a robust mathematical definition of privacy and a strong algorithmic framework allowing to design complex private algorithms

• DP induces a **privacy-utility trade-off** which depends on the **trust model**: the two extreme cases are the central (trusted curator) model and the local model (trust no one and nothing except oneself)

• In the context of **federated learning**, we can leverage appropriate **relaxations of local DP** to nearly **match the privacy-utility trade-off of the central model**
SOME OPEN PROBLEMS IN PRIVACY & ML

• Going beyond worst-case privacy-utility trade-offs: leverage the structure of some machine learning problems to design better DP algorithms

• Better privacy accounting: tight, automatic and personalized

• Correctness guarantees under malicious parties: make computation verifiable while preserving privacy guarantees

• Combining DP with secure multi-party computation: identify tractable secure primitives under which one can achieve trusted curator utility for many problems

• Concrete DP/FL deployments: match DP bounds to protection against specific attacks, articulate with the law (GDPR), make FL transparent to end-users
THANK YOU FOR YOUR ATTENTION!
QUESTIONS?


Rényi Differential Privacy.

In CSF.


In IEEE Symposium on Security and Privacy.


Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores.

In International Conference on Research in Computational Molecular Biology RECOMB.
Distributed Differentially Private Averaging with Improved Utility and Robustness to Malicious Parties.

Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes.
In ICML.

Membership Inference Attacks Against Machine Learning Models.
In IEEE Symposium on Security and Privacy (S&P).
RÉNYI DP

Definition (Rényi Differential Privacy)
Let $\alpha > 1$, $\epsilon > 0$. A randomized algorithm $\mathcal{A}$ is $(\alpha, \epsilon)$-RDP if for every adjacent datasets $\mathcal{D} \sim \mathcal{D}'$, we have:

$$D_\alpha (\mathcal{A}(\mathcal{D})\|\mathcal{A}(\mathcal{D}')) \leq \epsilon,$$

where $D_\alpha (P\|Q)$ is the Rényi divergence of order $\alpha$ between probability distributions $P$ and $Q$ defined as:

$$D_\alpha (P\|Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim Q} \left[ \frac{P(x)}{Q(x)} \right]^\alpha.$$

Proposition (From RDP to $(\epsilon, \delta)$-DP)
If $\mathcal{A}$ is an $(\alpha, \epsilon)$-RDP algorithm, then it also satisfies $(\epsilon + \frac{\log(1/\delta)}{\alpha-1}, \delta)$-DP for any $\delta \in (0, 1)$. 
Proposition (Gaussian mechanism in RDP)

Let $f$ be a function taking as input a dataset, and has $L2$ sensitivity bounded by $\Delta$. Then $A(D) = f(D) + \eta$ with $\eta \sim \mathcal{N}(0, \sigma^2 I)$ satisfies $(\alpha, \epsilon)$-RDP for any $\alpha > 1$ and $\epsilon = \frac{\alpha \Delta}{2\sigma^2}$.

Proposition (Composition under RDP)

If $A_1$ satisfies $(\alpha, \epsilon_1)$-RDP and $A_2$ satisfies $(\alpha, \epsilon_2)$-RDP, then $A = (A_1, A_2)$ satisfies $(\alpha, \epsilon_1 + \epsilon_2)$-RDP.

- RDP keeps tracks of the distribution of the privacy loss random variable
- Privacy accounting is done in RDP; then given the desired $\delta$ for the final guarantee, $\alpha$ is optimized (analytically or numerically) to get the best $\epsilon$
- In practice this is much better than resorting to advanced composition