Privacy Accounting and Quality Control in the Sage Differentially Private ML Platform

Mathias Lécuyer

With:
Riley Spahn, Kiran Vodrahalli, Roxana Geambasu, and Daniel Hsu
Machine Learning (ML) introduces a dangerous double standard for data protection

Example: messaging app
Example: messaging app

Traditional code

ML platform (e.g. TFX)

- auto-complete model
- ad targeting model
- recommendation model

Growing Database

messages, likes, clicks...
Example: messaging app

- User's messages (per access control restrictions)
- Traditional code
- Access control
- API
- ML platform (e.g. TFX)
  - Auto-complete model
  - Ad targeting model
  - Recommendation model

Growing Database

Messages, likes, clicks...
Example: messaging app

models and/or predictions
(based on everyone's messages, likes, clicks...)

Growing Database

Traditional code
Access control

API

ML platform (e.g. TFX)

auto-complete model
ad targeting model
recommendation model

messages, likes, clicks...
Example: messaging app

ML should only captures general trends from the data, but often captures **specific information about individual entries** in the dataset.
Example: messaging app

Language models over users’ emails leak secrets.
(Carlini+ '18)

messages, likes, clicks...
Growing Database

Traditional code
Access control

API
ML platform (e.g. TFX)

auto-complete model
ad targeting model
recommendation model
Example: messaging app

Membership in a training set can be inferred through prediction APIs. (Shokri+17)

Traditional code

Access control

API

ML platform (e.g. TFX)

auto-complete model

ad targeting model

recommendation model

Growing Database

messages, likes, clicks...

Recommenders leak information across users. [Calandrino'11]

Membership in a training set can be inferred through prediction APIs. (Shokri+17)
Example: messaging app

Recommenders leak information across users.  
(Calandrino'11)

ML platform (e.g. TFX)

- auto-complete model
- ad targeting model
- recommendation model

API

Traditional code

Access control

Growing Database

messages, likes, clicks...
Example: messaging app

- Making individual training algorithms Differentially Privacy (DP) is good but insufficient, because old data is reused many times.
- No system exists for managing multiple DP training algorithms to enforce a global DP guarantee.
Example: messaging app

• Making individual training algorithms Differentially Privacy (DP) is good but insufficient, because old data is reused many times.

• No system exists for managing multiple DP training algorithms to enforce a global DP guarantee.
Can we make Differential Privacy practical for ML applications?
Sage

• Enforces a global \((\varepsilon_g, \delta_g)\)-DP guarantee across all models ever released from a growing database.

• Tackles in practical ways two difficult DP challenges:
  1. “Running out of budget”
  2. “Privacy-utility tradeoff.”

![Diagram of Sage and ML platform (e.g. TFX)]
Outline

Motivation

Differential Privacy

Two practical challenges

Sage design

Evaluation
Differential Privacy (DP)

(Dwork+ '06)

• Developed to allow privacy-preserving statistical analyses on sensitive datasets (e.g., census, drug purchases, ...).

• First (and only) rigorous definition of privacy suitable for this use case.
• DP is a stability constraint on computations running on datasets: it requires that no single data point in an input dataset has a significant influence on the output.

• To achieve stability, randomness is added into the computation.
Definition

• DP is a stability constraint on computations running on datasets: it requires that no single data point in an input dataset has a significant influence on the output.

• To achieve stability, randomness is added into the computation.

• A randomized computation \( f: D \to O \), is \((\varepsilon, \delta)\)-DP if for any pair of datasets \( D \) and \( D' \) differing in one entry, and for any output set \( S \subseteq O \):

\[
P(f(D) \in S) \leq e^\varepsilon P(f(D') \in S) + \delta
\]
DP in ML

• Approach: make training algorithms DP.

• It prevents membership query and reconstruction attacks (Steinke-Ullman '14; Dwork+ '15; Carlini+ '18).

• DP versions exist for most ML training algorithms:
  - Stochastic gradient descent (SGD) (Abadi+16, Yu+19).
  - Various regressions (Chaudhuri+08, Kifer+12, Nikolaenko+13, Talwar+15).
  - Collaborative filtering (McSherry+09).
  - Language models (McMahan+18).
  - Feature and model selection (Chaudhuri+13, Smith+13).
  - Model evaluation (Boyd+15).
  - Tensorflow/privacy implements several of these algorithms (McMahan+19).
Outline

Motivation

Differential Privacy

Two practical challenges

Sage design

Evaluation
Challenge 1 - Running out of privacy budget

Most DP work focuses on a fixed database model:

• Each model consumes some privacy budget.
• When the budget is exhausted, the data cannot be used anymore: the system can "run out of budget".
Challenge 1 - Running out of privacy budget

Most DP work focuses on a fixed database model:

- Each model consumes some privacy budget.
- When the budget is exhausted, the data cannot be used anymore: the system can "run out of budget".
Most DP work focuses on a fixed database model:

- Each model consumes some privacy budget.
- When the budget is exhausted, the data cannot be used anymore: the system can "run out of budget".

**Challenge 1 - Running out of privacy budget**
Most DP work focuses on a fixed database model:

- Each model consumes some privacy budget.
- When the budget is exhausted, the data cannot be used anymore: the system can "run out of budget".
Challenge 2 - Privacy/utility trade-off
Challenge 2 - Privacy/utility trade-off

More utility

More privacy

Linear Regression

Deep Neural Network
Outline

Motivation

Differential Privacy

Two practical challenges

Sage design

Evaluation
Key realization: ML platforms operate on a growing database.
Interaction model:

- Split the growing database into time based blocks.
- Models can adaptively combine blocks to form larger datasets.
- Account for privacy loss only against blocks used by each model.
- Models can influence future data and privacy budgets.
Interaction model:
• Split the growing database into time based blocks.
• Models can adaptively combine blocks to form larger datasets.
• Account for privacy loss only against blocks used by each model.
• Models can influence future data and privacy budgets.
Sage block composition (challenge 1)

Interaction model:
- Split the growing database into time based blocks.
- Models can adaptively combine blocks to form larger datasets.
- Account for privacy loss only against blocks used by each model.
- Models can influence future data and privacy budgets.
Sage block composition (challenge 1)

Theorem:

\[ | \text{PrivacyLoss(stream)} | \leq \max_k | \text{PrivacyLoss}(D_k) | \]
Why is this important?

- Controlling each block’s privacy loss controls the global privacy loss.
- New blocks arrive with zero loss and constantly renew the budget.

Theorem:

\[ | \text{PrivacyLoss(stream)} | \leq \max_k | \text{PrivacyLoss}(D_k) | \]
Iterative training (challenge 2)
Iterative training (challenge 2)

- ML platform (e.g. TFX)
- (ε, δ)-DP ML model
- Sage access control global (ε_g, δ_g)-DP
- Privacy loss
- More utility
- More data
- More privacy
- MSE (x10^-3)
  - Non DP
  - DP (ε=1.0)
  - DP (ε=0.1)
- Training Samples
  - 10,000
  - 1,000,000
  - 100,000,000
- Good?
Iterative training (challenge 2)

- Adaptively trains on growing data and/or privacy budgets.
- Release when w.h.p. model accuracy surpasses a target.
- Accounts for the impact of DP noise in TFX-evaluate to give high-probability assessment of model accuracy.
Iterative training (challenge 2)

- Adaptively trains on growing data and/or privacy budgets.
- Release when w.h.p. model accuracy surpasses a target.
- Accounts for the impact of DP noise in TFX-evaluate to give high-probability assessment of model accuracy.
Iterative training (challenge 2)

- Adaptively trains on growing data and/or privacy budgets.
- Release when w.h.p. model accuracy surpasses a target.
- Accounts for the impact of DP noise in TFX-evaluate to give high-probability assessment of model accuracy.

Statistical test for evaluation:
\[ P(\text{acc} < \tau) \leq \eta \] over sampling of test set.
Iterative training (challenge 2)

- Adaptsively trains on growing data and/or privacy budgets.
- Release when w.h.p. model accuracy surpasses a target.
- Accounts for the impact of DP noise in TFX-evaluate to give high-probability assessment of model accuracy.

Statistical test for evaluation:

\[ P(\text{acc} < \tau) \leq \eta \text{ over sampling of test set and DP noise.} \]

\[
\begin{align*}
\overline{\mathcal{L}}_{\text{te}}^{dp}(f^{dp}) &+ \sqrt{\frac{2B\overline{\mathcal{L}}_{\text{te}}^{dp}(f^{dp})\ln(3/\eta)}{n_{\text{te}}^{dp}}} + \frac{4B\ln(3/\eta)}{n_{\text{te}}^{dp}} \leq \tau_{\text{loss}}
\end{align*}
\]
Sage Architecture

ML platform (e.g. TFX)

(ε, δ)-DP ML model

(ε/2, δ/2)-DP model training
(ε/2, δ/2)-DP model validation

dataset, ε, δ

request dataset

retry

release

reject/timeout

Models meet quality goal w.h.p

Traditional access control

Privacy loss

Sage access control

global (εg, δg)-DP

Time

Models meet quality goal w.h.p
Outline

Motivation

Differential Privacy

Two practical challenges

Sage design

Evaluation
Evaluation:

2. Importance of iterative training and DP aware performance tests.
3. Continuous operation on multiple models and growing database.
1. Benefits of block composition versus traditional DP composition

Data points used to reach target

Required sample size

Better model

MSE Target (x10^-3)

- Traditional DP composition
- Sage
2. Importance of iterative training and DP aware performance tests

<table>
<thead>
<tr>
<th>Test methodology</th>
<th>Non DP</th>
<th>DP + UB</th>
<th>Sage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure rate at 1% proba.</td>
<td>0.2%</td>
<td>1.7%</td>
<td>0.3%</td>
</tr>
</tbody>
</table>
3. Continuous operation on multiple models and growing database
Summary

• DP literature has mostly focused on individual ML algorithms running on static databases (which don’t incorporate new data).

• ML workloads operate on growing databases: models incorporate new data and (adaptively) reuse old data.

• Sage is the first to adapt DP theory and practice to ML workloads on growing databases, for data protection.
  - Opens an exciting design space for efficient privacy resource allocation!