DP-Finder: Finding Differential Privacy Violations by Sampling and Optimization

Benjamin Bichsel  Timon Gehr  Dana Drachsler-Cohen  Petar Tsankov  Martin Vechev

SRILAB  ETH Zürich
Differential Privacy – Basic Setting
Differential Privacy – Basic Setting

What about my privacy?
Differential Privacy - Intuition

Change my data

- # disease + noise

7.3

or

7.6
Differential Privacy – More Abstractly

Neighboring $x$ $x'$

Attacker check $F(x) \in \Phi$?

Attacker check $F(x') \in \Phi$?
Differential Privacy - Definition

\[ \Pr[F(x) \in \Phi] \leq \exp(\varepsilon) \approx 1 + \varepsilon \]

Challenges induced by DP:

- Proving/checking \( \varepsilon \)-DP is hard (buggy algorithms)
- Proof strategies not complete
- Proofs only provide upper bounds
\(\varepsilon\)-DP Counterexamples

that violate \(\varepsilon\)-DP:

\[
\frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]} > \exp(\varepsilon)
\]

\[
\Leftrightarrow
\]

\[
\log \frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]} > \varepsilon
\]
\( \varepsilon \)-DP Counterexamples

that violate \( \varepsilon \)-DP:

\[
\frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]} > \exp(\varepsilon)
\]

\[
\iff \log \frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]} > \varepsilon
\]

Maximize \( \varepsilon(x, x', \Phi) \)
Bounds on "true" $\varepsilon$

**Evaluation:** We get precise and large $\varepsilon$, close to known upper bounds.

- Counterexample: 5%-DP
- Counterexample: 9.9%-DP
- Counterexample: 15%-DP

Proven: 10%-DP ($\varepsilon = 10\% = 0.1$)
**ε-DP Counterexamples**

**Goal:** Maximize $\varepsilon(x, x', \Phi)$

**Challenge 1:** Expensive to compute $\varepsilon$ precisely

**Challenge 2:** Search space is sparse: Few $x, x', \Phi$ lead to large $\varepsilon(x, x', \Phi)$

Estimate $\varepsilon$ by sampling

Make $\hat{\varepsilon}$ differentiable
Step 1: Estimate $\varepsilon$
Estimating $\varepsilon$

$$
\varepsilon(x, x', \Phi) := \log \frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]}
$$
Estimating $\varepsilon$

$$\varepsilon(x, x', \Phi) := \log \frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]}$$

Estimate

$$\hat{\Pr}[F(x) \in \Phi] = \frac{1}{n} \sum_{i=1}^{n} \text{check}^i_{F,\Phi}(x)$$

$x$

$F(x)$

$\text{check}^i_{F,\Phi}(x)$

<table>
<thead>
<tr>
<th>$x$</th>
<th>$F(x)$</th>
<th>$\text{check}^i_{F,\Phi}(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$7.3$</td>
<td>yes</td>
<td>$33%$</td>
</tr>
<tr>
<td>$7.6$</td>
<td>no</td>
<td>$33%$</td>
</tr>
<tr>
<td>$6.8$</td>
<td>yes</td>
<td>$67%$</td>
</tr>
</tbody>
</table>
How precise is our estimate?

Counterexample: 9.9% ± 10%-DP

VS

Counterexample: 9.9% ± 2 · 10^{-3}-DP

Precision of $\varepsilon$

Precision of $\Pr[F(x) \in \Phi]$ and $\Pr[F(x') \in \Phi]$

Sampling effort $n$

Exponential search
Estimating precisely is expensive

Estimating $\varepsilon$ up to an error of $2 \cdot 10^{-3}$ with confidence of 90%
Applying the M-CLT (Correlation)

\[
\frac{1}{n} \sum_{i=1}^{n} \text{check}_{F,\Phi}^i(x)
\]

Follows 2D Gaussian distribution

\[
\frac{1}{n} \sum_{i=1}^{n} \text{check}_{F,\Phi}^i(x')
\]
Obtaining a Confidence Interval for $\varepsilon$

Joint likelihood of
\[
\left( \Pr\left[ F(x) \in \Phi \right] \right) / \left( \Pr\left[ F(x') \in \Phi \right] \right)
\]

Likelihood of $\varepsilon(x, x', \Phi)$

Confidence Interval for $\varepsilon(x, x', \Phi)$

Distribution of Gauss (correlated):

How precise is our estimate?

Counterexample: 
9.9% ± 10% - DP

VS

Counterexample: 
9.9% ± 2 \cdot 10^{-3} - DP

![Graph showing sampling effort vs. probability of \( F(x) \in \Phi \).]
Step 2: Finding Counterexamples

Make $\hat{\xi}$ differentiable
How can we optimize our estimate?

\[ \hat{\epsilon}(x, x', \Phi) = \log \frac{1}{n} \sum_{i=1}^{n} \text{check}_{F, \Phi}^i (x') \]

maximize

Goals

- Make differentiable
- Preserve semantics

\[ \neg B \sim 1 - B \]

\[ B_1 \land B_2 \sim B_1 \cdot B_2 \]

if \( B \): \( \{ x = E_1 \} \) else : \( \{ x = E_2 \} \sim x = B \cdot E_1 + (1 - B) \cdot E_2 \]
How can we optimize our estimate?

\[
\hat{e}(x, x', \Phi) = \log \frac{1}{n} \sum_{i=1}^{n} \text{check}_{F,\Phi}^i (x')
\]

Maximize using SLSQP (supports hard constraints for neighborhood)

Random starting point (+ restart)

What about division by zero?

What about very small denominators?
Main differences to Ding et al.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Ding et al.</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem statement</td>
<td>$\varepsilon(x, x', \Phi) &gt; \varepsilon_0$?</td>
<td>Maximize $\varepsilon(x, x', \Phi)$</td>
</tr>
<tr>
<td>Approach</td>
<td>Statistical tests</td>
<td>Estimate + confidence interval</td>
</tr>
<tr>
<td>Search</td>
<td>By patterns</td>
<td>Gradient descent (incremental)</td>
</tr>
</tbody>
</table>
Evaluation

- How **precise** is the differentiable estimate?
- How **efficient** is DP-Finder in finding violations compared to random search?
Precision of Differentiable Estimate

![Graph showing precision of differentiable estimate for various algorithms: AT1, AT2, AT3, AT4, AT5, expMech, noisyMax, sum. The x-axis represents the algorithms, and the y-axis shows the precision values.
Random vs Optimized

[Graph showing comparisons between random and optimized start processes]

Random start vs Optimized
Conclusion

Differential Privacy

Estimate $\varepsilon$

Finding Counterexamples

$\varepsilon$-DP Counterexamples