Bounding User Contributions: A Bias-Variance Trade-off in Differential Privacy

Kareem Amin, Alex Kulesza, Andrés Muñoz Medina, Sergei Vassilvitskii
Google Research NY
Typical DP assumption:
One user = one example

Reality:
Users contribute many times
High cap = excessive noise

Low cap = biased data

We investigate this bias-variance trade-off using tools from learning theory
Setting

Infinite collection of users

- Distribution $P$ over users
- Each user has a unique distribution over examples
- i.i.d. data: first sample a user from $P$, then sample the user’s distribution
Learning

- Cap each user at a $\pi_0$ fraction of the dataset
- Run a standard differentially private ERM algorithm
Result

\[ \mathcal{L}(h_{\text{priv}}) \leq \inf_{h \in H} \mathcal{L}(h) + \text{Bias due to capping} + \text{Finite sample variance} + \text{Privacy noise variance} \]
Result

\[ \mathcal{L}(h_{\text{priv}}) \leq \inf_{h \in H} \mathcal{L}(h) + O \left( \sqrt{\frac{\text{Var}(H)}{\tau_0}} \right) + \text{Finite sample variance} + \text{Privacy noise variance} \]
Result

\[ \mathcal{L}(h_{\text{priv}}) \leq \inf_{h \in H} \mathcal{L}(h) + O \left( \sqrt{\frac{\text{Var}(H)}{\tau_0}} \right) + \tilde{O} \left( \sqrt{\frac{1}{\tau_0 n}} \right) + \text{Privacy noise variance} \]

- Bias due to capping
- Finite sample variance
Result

\[
\mathcal{L}(h_{\text{priv}}) \leq \inf_{h \in H} \mathcal{L}(h) + O \left( \sqrt{\frac{\text{Var}(H)}{\tau_0}} \right) + \tilde{O} \left( \sqrt{\frac{1}{\tau_0 n}} \right) + O \left( \frac{1}{K^2(\tau_0)} \right)
\]

- Bias due to capping
- Finite sample variance
- Privacy noise variance
The Cost of Privacy

As $n \rightarrow \infty$...

$$\mathcal{L}(h_{\text{priv}}) \leq \inf_{h \in H} \mathcal{L}(h) + O \left( \sqrt{\frac{\text{Var}(H)}{\tau_0}} \right) + \tilde{O} \left( \sqrt{\frac{1}{\tau_0 n}} \right) + O \left( \frac{1}{K^2(\tau_0)} \right)$$
The Cost of Privacy

As \( n \to \infty \) ...

\[
\mathcal{L}(h_{\text{priv}}) \leq \inf_{h \in H} \mathcal{L}(h) + O \left( \sqrt{\frac{\text{Var}(H)}{\tau_0}} \right) + \tilde{O} \left( \sqrt{\frac{1}{\tau_0 n}} \right) + O \left( \frac{1}{K^2(\tau_0)} \right)
\]

For privacy noise to vanish, \( \tau_0 \to 0 \)
The Cost of Privacy

As $n \to \infty$ ...

\[ \mathcal{L}(h_{\text{priv}}) \leq \inf_{h \in H} \mathcal{L}(h) + O \left( \sqrt{\frac{\text{Var}(H)}{\tau_0}} \right) + \tilde{O} \left( \sqrt{\frac{1}{\tau_0 n}} \right) + O \left( \frac{1}{K^2(\tau_0)} \right) \]

But then bias grows without bound

For privacy noise to vanish, $\tau_0 \to 0$
The Cost of Privacy

As \( n \to \infty \) …

\[
\mathcal{L}(h_{\text{priv}}) \leq \inf_{h \in H} \mathcal{L}(h) + O \left( \sqrt{\frac{\text{Var}(H)}{\tau_0}} \right) + \tilde{O} \left( \sqrt{\frac{1}{\tau_0 n}} \right) + O \left( \frac{1}{K^2(\tau_0)} \right)
\]

But then bias grows without bound

For privacy noise to vanish, \( \tau_0 \to 0 \)

Privacy incurs a fixed cost: we cannot recover optimal error even when \( n \to \infty \)