Differential Privacy for Machine Learning

Master IASD, Université PSL

February 2024

Recommended Readings

Rererences used for this lecture:

- Deep Learning with Differential Privacy – Abadi et al. ACM CCS 2016.
- Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data – Papernot et al. ICLR 2017
- Model-Agnostic Private Learning via Stability – Bassily et al. NeurIPS 2018.

Modern ML

- Huge number of data points: $n \sim 1,000,000$ (Example: ImageNet has 14 million images!)
- Huge number of model parameters: $p \sim 1,000,000$ (Example: Resnet-18 has 11 million parameters!)
- Loss function typically not convex, not very smooth.

Table of Contents

SGD vs GD

- Computing the full gradient is $O(n)$ - very expensive.
- Instead, gradient is typically computed on a small batch of data points ($m \sim 100$) called mini-batch.
- In each iteration, m out of n data points are randomly sampled to form a mini-batch. Then, model parameters are updated with the mini-batch gradient.
- SGD with a small batch-size typically takes more iterations for convergence than the one with large batch-size.

DP-SGD

Algorithm:
The algorithm is identical to DP-GD except that we use mini-batch gradient in place of full-gradient. Importantly, the gradient is clipped at some max value in every iteration.

$$g_t \leftarrow \nabla \mathcal{L}(\theta_t; \mathcal{D}_t)$$
$$\widetilde{g}_t \leftarrow \min \{C, g_t / \|g_t\|_2\}$$
$$\theta_{t+1} \leftarrow \theta_t - \eta \widetilde{g}_t + \mathbb{N}(0, \sigma^2 \mathcal{I}_p)$$

DP-SGD

Utility:

- The mini-batch gradient is equal to the full gradient in expectation.
- Random subsampling for mini-batch introduces additional variance term in utility analysis.
- The effect of gradient-clipping is not fully understood.

Privacy:

- Sensitivity of the gradient is deliberately bounded by clipping.
- Two new techniques for tighter privacy analysis.
  1. Privacy amplification by sub-sampling
  2. Moments accountant

Privacy amplification by sub-sampling

Theorem

Let $M(\mathcal{D})$ be an ε-DP mechanism. Then the mechanism $M'(\mathcal{D})$ that outputs the result of M on a random subsample $\mathcal{D}_\gamma$ of $\mathcal{D}$ of size γn is $2\gamma(e^\varepsilon - e^{-\varepsilon})$-DP.

Proof Sketch: Let $\mathcal{D}$ and $\mathcal{D}'$ be two neighboring datasets of size n, differing in d_i. Let $S \subset [n]$ be the subset of indices sampled. Let $R \subset [n]\backslash\{i\}$. For any event E,

$$\frac{\mathbb{P}(M'(\mathcal{D}) \in E)}{\mathbb{P}(M'(\mathcal{D}') \in E)} = \frac{\gamma\mathbb{P}(M'(\mathcal{D}) \in E|i \in S) + (1-\gamma)\mathbb{P}(M'(\mathcal{D}) \in E|i \notin S)}{\gamma\mathbb{P}(M'(\mathcal{D}') \in E|i \in S) + (1-\gamma)\mathbb{P}(M'(\mathcal{D}') \in E|i \notin S)}$$

$$= \frac{\mathbb{E}_{R,j\neq i}\left[\gamma\mathbb{P}(M'(\mathcal{D}) \in E|S = R\cup\{i\}) + (1-\gamma)\mathbb{P}(M'(\mathcal{D}) \in E|S = R\cup\{j\}, j\neq i)\right]}{\mathbb{E}_{R,j\neq i}\left[\gamma\mathbb{P}(M'(\mathcal{D}') \in E|S = R\cup\{i\}) + (1-\gamma)\mathbb{P}(M'(\mathcal{D}') \in E|S = R\cup\{j\}, j\neq i)\right]}$$

$$\leq \frac{(\gamma e^\varepsilon + 1 - \gamma)\mathbb{E}_{R,j\neq i}\mathbb{P}(M'(\mathcal{D}) \in E|S = R\cup\{j\}, j\neq i)}{(\gamma e^{-\varepsilon} + 1 - \gamma)\mathbb{E}_{R,j\neq i}\mathbb{P}(M'(\mathcal{D}) \in E|S = R\cup\{j\}, j\neq i)}$$

$$= \frac{(\gamma e^\varepsilon + 1 - \gamma)}{(\gamma e^{-\varepsilon} + 1 - \gamma)} = \frac{1 + \gamma(e^\varepsilon - 1)}{1 + \gamma(e^{-\varepsilon} - 1)}.$$

Privacy amplification by sub-sampling

Proof Sketch: (continued)
Therefore, M' is ε'-DP where,

$$
\begin{aligned}
\varepsilon' &\le \log\left(\frac{1+\gamma(e^{\varepsilon}-1)}{1+\gamma(e^{-\varepsilon}-1)}\right) \\
&\le 2\gamma(e^{\varepsilon}-e^{-\varepsilon}).
\end{aligned}
$$

$\square$

- For small ε, $2\gamma(e^{\varepsilon}-e^{-\varepsilon}) \approx 4\gamma\varepsilon$. Hence, subsampling by γ fraction amplifies privacy by a factor of 4γ!
- $\gamma \ll 1$ is better for privacy, but worse for utility.
- DP-SGD with batch-size m gets a privacy amplification of $O(m/n)$.

Moments Accountant

Theorem (Advanced Composition Theorem (Dwork & Roth))

For all $\varepsilon, \delta, \delta' \geq 0$, the class of (ε, δ)-DP mechanisms satisfies $(\varepsilon', k\delta + \delta')$-DP under k-fold adaptive composition for

$$\varepsilon' = \sqrt{2k \ln(1/\delta')}\varepsilon + k\varepsilon(e^{\varepsilon} - 1).$$

- For small ε, advanced composition gives $(O(\sqrt{k}\varepsilon), O(k\delta))$-DP for the composition of k (ε, δ)-DP mechanisms.
- Advanced composition theory is independent of the specifics of each mechanism in compositino.

Moments Accountant

Recall that the **privacy loss random variable** for a mechanism M is defined as,

$$\ell(y; M, \mathcal{D}, \mathcal{D}') := \log \frac{\mathbb{P}(M(\mathcal{D}) = y)}{\mathbb{P}(M(\mathcal{D}') = y)}.$$

Observe that M is ε-DP if $\ell(y; M, \mathcal{D}, \mathcal{D}') \leq \varepsilon$ for all y.
Define the **moments accountant** parametrized by $\lambda > 0$ as,

$$\alpha_M(\lambda) := \max_{\mathcal{D}, \mathcal{D}' : d(\mathcal{D}, \mathcal{D}') = 1} \log \mathbb{E}[\exp(\lambda \ell(y; M, \mathcal{D}, \mathcal{D}'))].$$

Moments Accountant

Theorem

For any $\varepsilon > 0$, a mechanism M with moments accountant $\alpha_M(\lambda)$ is (ε, δ)-DP for,

$$\delta = \min_{\lambda > 0} \exp(\alpha_M(\lambda) - \lambda\varepsilon).$$

Proof Sketch:

$$\begin{aligned}\mathbb{P}(\ell(y; M, \mathcal{D}, \mathcal{D}') > \varepsilon) &= \mathbb{P}(\exp(\lambda\ell(y; M, \mathcal{D}, \mathcal{D}')) > \exp(\lambda\varepsilon)) \\ &\leq \frac{\mathbb{E}[\exp(\lambda\ell(y; M, \mathcal{D}, \mathcal{D}'))]}{\exp(\lambda\varepsilon)} = \exp(\alpha_M(\lambda) - \lambda\varepsilon).\end{aligned}$$

Let bad event $B = \{y : \ell(y; M, \mathcal{D}, \mathcal{D}') \geq \varepsilon\}$. For any event E,

$$\begin{aligned}\mathbb{P}(M(\mathcal{D}) \in E) &= \mathbb{P}(M(\mathcal{D}) \in E \cap B^c) + \mathbb{P}(M(\mathcal{D}) \in E \cap B) \\ &\leq e^{\varepsilon}\mathbb{P}(M(\mathcal{D}') \in E \cap B^c) + \mathbb{P}(M(\mathcal{D}) \in B) \\ &\leq e^{\varepsilon}\mathbb{P}(M(\mathcal{D}') \in E) + \exp(\alpha_M(\lambda) - \lambda\varepsilon). \qquad \square\end{aligned}$$

Moments Accountant

Composition

Let a mechanism M be a composition of mechanisms $M_1, \ldots, M_k$. Then for any $\lambda > 0$,

$$\alpha_M(\lambda) \leq \sum_{i=1}^{k} \alpha_{M_i}(\lambda).$$

Moments Accountant

Using moments accountant with DP-SGD:

- In each iteration $t \in \{1, \ldots, T\}$ of DP-SGD, we have M_t, a Gaussian mechanism that is amplified by sub-sampling. Hence, $\alpha_{M_t}(\lambda)$ can be computed for a range of $\lambda \in (0, \lambda_{max})$ using numerical integration.
  - Example: Condider $1D$. Without loss of generality, assume $M(\mathcal{D}) \sim \mathbb{N}(0, \sigma^2) = \mu_0$ and $M(\mathcal{D}') \sim \mathbb{N}(1, \sigma^2) = \mu_1$. Let $\mu = (1-\gamma)\mu_0 + \gamma\mu_1$. Then, $\alpha = \log \max\{E_1, E_2\}$ where,

$$E_1 = \mathbb{E}_{z \sim \mu_0}[(\mu_0(z)/\mu(z))^\lambda],$$
$$E_2 = \mathbb{E}_{z \sim \mu}[(\mu(z)/\mu_0(z))^\lambda].$$

- For T iterations, accumulate the moments accountants and then compute $\alpha_M(\lambda) \leq \sum_{t=1}^{T} \alpha_{M_t}(\lambda)$.
- Finally, choose optimal $\lambda > 0$ to find best possible (ε, δ) using the formula $\delta = \min_{\lambda > 0} \exp(\alpha_M(\lambda) - \lambda\varepsilon)$.

Moments Accountant

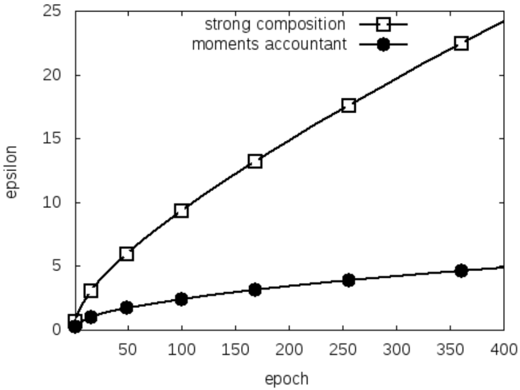

Figure: Comparison of privacy guarantee with advanced composition vs moments accountant for $\gamma = 0.01$, $\delta = 10^{-5}$ and $\sigma = 4$. Figure taken from Abadi et al. 2016.

Table of Contents

PATE - Private Aggregation of Teacher Ensembles

Algorithm:

1. Train an ensemble of teachers on disjoint subsets of sensitive data.
2. Train a student model on public unlabeled data labeled using the teacher-ensemble (semi-supervised learning).
3. For private inference, use the student model.

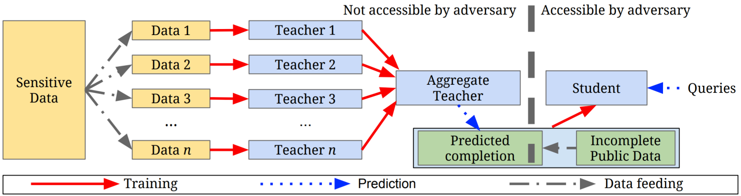

Figure: PATE Block Diagram

PATE

Utility:

- If each teacher gets access to a large-enough fraction of the dataset (i.e., if the number of teachers is small enough), then each teacher can be trained to high accuracy. Subsequently, aggregate teacher has high accuracy.
- Subsampling and aggregation is similar to boosting in ML, and may help accuracy.
- Availability of public unlabeled data may help improve accuracy if semi-supervised learning methods are used.

PATE

Privacy: Per-query privacy of the teacher ensemble

- Simple PATE: Noisy arg-max via Laplace mechanism - gives ε-DP.
  - Let $\{f_i\}_{i=1}^{n_T}$ denote the classifiers for each of the n_T teachers. For each label $j \in [L]$, $c_j(\mathcal{D}) = |\{i \in [n_T] : f_i(\mathcal{D}_i) = j\}|$ is the count for label j.

$$\hat{f}(\mathcal{D}) = \arg\max_{j \in [L]} \{c_j(\mathcal{D}) + Lap(0, 1/\varepsilon)\}.$$

- Improved PATE: Subsampling and aggregation via stability - gives (ε, δ)-DP.
  - Intuition: If there is a clear majority among the teachers, then we can release the exact arg-max without losing privacy.

PATE

Privacy: Overall privacy of the student model

- Simple PATE: Advanced composition and post-processing
  - If the aggregation mechanism is ε-DP, then the training data to the student model with k labelings of the teacher ensemble is $O(\sqrt{k}\varepsilon)$-DP by advanced composition.
  - The output of the student model (irrespective of the number of times it is queried) is also $O(\sqrt{k}\varepsilon)$-DP by post-processing property.
  - More teachers is better for privacy, but worse for utility.
- Improved PATE: Sparse vector technique
  - Intuition: By refusing to answer unstable queries, one can get a better privacy guarantee than strong composition for a large number of queries. Similar to "Above Threshold" mechanism, we have stability threshold.

PATE

Privacy: Per-query privacy of the teacher ensemble

- Simple PATE: Noisy arg-max via Laplace mechanism - gives ε-DP.
  - Let $\{f_i\}_{i=1}^{n_T}$ denote the classifiers for each of the n_T teachers. For each label $j \in [L]$, $c_j(\mathcal{D}) = |\{i \in [n_T] : f_i(\mathcal{D}_i) = j\}|$ is the count for label j.

$$\hat{f}(\mathcal{D}) = \arg\max_{j \in [L]} \{c_j(\mathcal{D}) + Lap(0, 1/\varepsilon)\}.$$

- Improved PATE: Subsampling and aggregation via stability - gives (ε, δ)-DP.
  - Intuition: If there is a clear majority among the teachers, then we can release the exact arg-max without losing privacy.

Notions of Stability

Let f be a function on dataset $\mathcal{D}$.

- **Subsampling Stability:** f is γ-subsampling stable on $\mathcal{D}$ if $f(\mathcal{D}_\gamma) = f(\mathcal{D})$ with probability at least $3/4$ when $\mathcal{D}_\gamma$ is a random subsample from $\mathcal{D}$ which includes each entry independently with probability γ.
- **Perturbation Stability:** f is k-perturbation stable on $\mathcal{D}$ if $f(\mathcal{D}') = f(\mathcal{D})$ for all $\mathcal{D}'$ such that $d(\mathcal{D}, \mathcal{D}') \leq k$. We say that f is *stable* on $\mathcal{D}$ if it is at least 1-stable on $\mathcal{D}$, and *unstable* otherwise.

Define **distance to instability** of a dataset $\mathcal{D}$ with respect to a function f as follows.

$$dist_f(\mathcal{D}) = \arg\max\{k \in [n] : f(\mathcal{D})\ is\ k - perturbation\ stable\}.$$

Privacy via Stability

M_{stab} Privacy via Stability

Input: Dataset $\mathcal{D}$, function f, threshold T privacy budget (ε, δ).

1. $T \leftarrow \frac{\log(1/\delta)}{\varepsilon}$.
2. $\hat{d} \leftarrow dist_f(\mathcal{D}) + Lap(0, 1/\varepsilon)$.
3. If $\hat{d} > T$, then $\hat{f} = f(\mathcal{D})$, else $\hat{f} = fail$.

Output: $\hat{f}$.

Privacy via Stability

- M_{stab} returns the *exact* output on stable datasets!
- Computing $dist_f(\mathcal{D})$ can be very intensive.
- M_{stab} is a special case of a much broader class of mechanisms that fall under Propose-Test-Release (PTR) framework.

Privacy of M_{stab}

Theorem

M_{stab} is (ε, δ)-DP.

Proof sketch: Take any two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$.

- **Case I:** $f(\mathcal{D}) = f(\mathcal{D}')$. Observe that the sensitivity of $dist_f$ is 1. Hence, by Laplace mechanism, $\hat{d}$ is an ε-DP estimate for $dist_f(\mathcal{D})$. Hence, by post-processing property, M_{stab} is ε-DP.
- **Case II:** $f(\mathcal{D}) \neq f(\mathcal{D}')$. In this case, $dist_f(\mathcal{D}) = dist_f(\mathcal{D}') = 0$. Hence, $\hat{d}$ is a zero-mean Laplace random variable. By a tail bound, $\hat{d} \leq T$ w.p. at least $1 - \delta$. Hence, M_{stab} returns the same output i.e. $fail$ on both datasets w.p. at least $1 - \delta$. Therefore, M_{stab} is ε-DP w.p. at least $1 - \delta$.

□

Utility of M_{stab}

Theorem

If any $\beta > 0$, if f is $\frac{\log(1/\delta)+\log(1/\beta)}{\varepsilon}$-perturbation stable on $\mathcal{D}$, then $M_{stab}(\mathcal{D}) = f(\mathcal{D})$ with probability at least $1-\beta$.

Proof sketch: If $dist_f(\mathcal{D}) > \frac{\log(1/\delta)+\log(1/\beta)}{\varepsilon}$, then by a tail bound on Laplace distribution, we get $\hat{d} > \frac{\log(1/\delta)}{\varepsilon}$ with probability at least $1-\beta$. Hence, with probability at least $1-\beta$, $M_{stab}(\mathcal{D}) = f(\mathcal{D})$.

$\square$

Privacy via Subsample and Aggregate

M_{samp} Privacy via Subsample and Aggregate

Input: Dataset $\mathcal{D}$, function f, privacy budget (ε, δ).

1. $\gamma \leftarrow \frac{\varepsilon}{32\log(1/\delta)}$, $n_T \leftarrow \frac{\log(n/\delta)}{q^2}$.
2. Subsample n_T datasets $\mathcal{D}_1, \ldots, \mathcal{D}_{n_T}$ where $\mathcal{D}_i$ includes each $d \in \mathcal{D}$ w.p.γ.
3. If some $d \in \mathcal{D}$ appears in more than $2\gamma n_T$ subsampled datasets, then $\hat{f} = fail$.
4. Else,
    1. For each possible output y of f, $count(y) \leftarrow |\{i : f(\mathcal{D}_i) = y\}|$.
    2. $\hat{d} \leftarrow \frac{count_{(1)} - count_{(2)}}{4\gamma n_T} - 1 + Lap(0, 1/\varepsilon)$.
    3. If $\hat{d} > \frac{\log(1/\delta)}{\varepsilon}$, then $\hat{f} = \arg\max_y count(y)$, else $\hat{f} = fail$.

Output: $\hat{f}$.

Privacy of M_{samp}

Theorem

M_{samp} is (ε, δ)-DP.

Proof sketch:
Let Z_{ij} be a bernoulli random variable indicating the event $d_i \in \mathcal{D}_j$ where $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, n_T\}$. $\mathbb{E}[Z_{ij}] = \gamma$. From Hoeffding inequality,

$$\mathbb{P}(\sum_{j=1}^{n_T} Z_{ij} > 2\gamma n_T) \leq e^{-2(\gamma n_T)^2/n_T}.$$

Hence, the bad event B of some $d \in \mathcal{D}$ appearing in more than $2\gamma n_T$ subsampled datasets occurs with probability at most δ.
Conditioned on B^c, observe that the counts $count_{(1)}, count_{(2)}$ can change by at most $2\gamma n_T$ by changing one data point. Hence, $\frac{count_{(1)} - count_{(2)}}{4\gamma n_T}$ has sensitivity of 1, implying that $\hat{d}$ is ε-DP. The privacy guarantee now follows from that of M_{stab}.

Utility of M_{samp}

Theorem

If f is γ-subsampling stable on $\mathcal{D}$ for $\gamma = \frac{\varepsilon}{32\log(1/\delta)}$, then $M_{samp}(\mathcal{D}) = f(\mathcal{D})$ w.p. at least $1-3\delta$.

Proof sketch: Let Z_j be a bernoulli r.v. indicating the event $f(\mathcal{D}) = f(\mathcal{D}_i)$. From γ-subsampling stability of f, we have that $\mathbb{P}(Z_j = 1) = 3/4$ for all $j \in \{1, \ldots, n_T\}$. By Hoeffding inequality,

$$\mathbb{P}(\sum_j Z_j - 3n_T/4 < -n_T/8) \leq e^{-n_T/32}.$$

Hence, $count(f(\mathcal{D})) = \sum_j Z_j$ is at least $5n_T/8$ w.p. $1-\delta$. So, $count_{(1)} \geq 5n_T/8$ and $count_{(2)} \leq 3n_T/8$. Hence, $\frac{count_{(1)} - count_{(2)}}{4\gamma n_T} \geq \frac{1}{16\gamma}$.

Utility of M_{samp}

Proof sketch: (continued)

For $M_{samp}(\mathcal{D}) = f(\mathcal{D})$, we need $\hat{d} = \frac{count_{(1)} - count_{(2)}}{4\gamma n_T} - 1 + Lap(0, 1/\varepsilon) > \frac{\log(1/\delta)}{\varepsilon}$.
Using a tail bound on Laplace distribution, $Lap(0, 1/\varepsilon)$ does not go below $\frac{-\log(1/\delta)}{\varepsilon}$ w.p. $1 - \delta$. Hence, we need $\frac{1}{16\gamma} > \frac{2\log(1/\delta)}{\varepsilon}$.

$\square$

PATE

Privacy: Overall privacy of the student model

- Simple PATE: Advanced composition and post-processing
  - If the aggregation mechanism is ε-DP, then the training data to the student model with k labelings of the teacher ensemble is $O(\sqrt{k}\varepsilon)$-DP by advanced composition.
  - The output of the student model (irrespective of the number of times it is queried) is also $O(\sqrt{k}\varepsilon)$-DP by post-processing property.
  - More teachers is better for privacy, but worse for utility.
- Improved PATE: Sparse vector technique
  - Intuition: By refusing to answer unstable queries, one can get a better privacy guarantee than strong composition for a large number of queries. Similar to "Above Threshold" mechanism, we have stability threshold.

Recap: Above Threshold

Algorithm 1 Input is a private database D, an adaptively chosen stream of sensitivity 1 queries $f_1, \ldots$, and a threshold T. Output is a stream of responses $a_1, \ldots$

AboveThreshold$(D, \{f_i\}, T, \epsilon)$

- **Let** $\hat{T} = T + \mathrm{Lap}\left(\frac{2}{\epsilon}\right)$.
- **for** Each query i **do**
  - **Let** $\nu_i = \mathrm{Lap}(\frac{4}{\epsilon})$
  - **if** $f_i(D) + \nu_i \geq \hat{T}$ **then**
    - **Output** $a_i = \top$.
    - **Halt.**
  - **else**
    - **Output** $a_i = \bot$.
  - **end if**
- **end for**

Figure: Above Threshold Mechanism from Dwork & Roth

Sparse Vector Technique

Above Threshold Mechanism:

- Is ε-DP.
- Noise added to each query is $Lap(0, 4/\varepsilon)$.
- Halts after encountering the first query that exceeds the threshold.
- Outputs index of the last query.

Sparse Vector Mechanism:

- Is (ε, δ)-DP. (Apply advanced composition on Above Threshold mechanism.)
- Noise added to each query is $Lap(0, \sqrt{32Q \log(1/\delta)}\varepsilon)$.
- Halts after encountering Q queries that crosses the threshold.
- Outputs the indices of Q queries that exceed threshold.

Privacy via sparse vector technique

M_{svec} Privacy via sparse vector

Input: Dataset $\mathcal{D}$, function f, privacy budget (ε, δ), query set $\{f_1, \ldots, f_m\}$, unstable query count Q.

1. $q \leftarrow 0$, $\lambda \leftarrow \sqrt{32Q \log(1/\delta)}\varepsilon$, $T \leftarrow 2\lambda \log(2m/\delta)$.
2. $\hat{T} \leftarrow T + Lap(\lambda)$.
3. For $f \in \{f_1, \ldots, f_m\}$ and $c \leq Q$ do
    1. $\hat{f} \leftarrow M_{stab}(\mathcal{D}, f, T = \hat{T}, \varepsilon = 1/2\lambda)$.
    2. If $\hat{f} = fail$
        1. $c \leftarrow c + 1$.
        2. $\hat{T} \leftarrow T + Lap(\lambda)$.

Output: $\hat{f}$.

Privacy via sparse vector technique

- Noise added to each query is $O(\sqrt{Q}\log(m))$ as opposed to $O(\sqrt{m})$ with advanced composition.
- Allows for answering a lot more queries than with simple PATE.

Table of Contents

DP-SGD vs PATE

PATE	**DP-SGD**
Uses a group of "teacher" models to train a "student" model	Trains a single model using stochastic gradient descent
Model agnostic	Model specific
Privacy degrades with more queries on Teachers	Privacy degrades with more iterations of optimization
Techniques: stability, aggregation, sparse vector technique	Techniques: sub-sampling, moments accountant
Can work with heterogeneous data, heterogeneous models	Works with homogeneous data and model

Topics for further exploration

- Differentially private PAC learning
- Federated learning and Local differential privacy
- Connections with robust machine learning