DS603: Robust Machine Learning	Autumn 2025

Lecture 19 — Nov 5

Lecturer: Arjun Bhagoji *Scribe: Gourish Garg*

19.1 Differential privacy

19.1.1 Motivation

Neural networks often memorize unique examples during training. This makes models trained on sensitive user information such as medical records are vulnerable to information leakage. [1] A trivial solution one might propose is to edit or remove identifying information from the training data so that specific indentifiers cannot be deduced. This approach is formally known as *anonymization*. However, anonymization implicitly assumes the non-existence of any publicly available auxiliary dataset or model that could help re-identify identifiers. [2] shows that even after anonymization, publicly available combinations of key attributes can uniquely identify individuals with high probability.

Therefore, we need a more rigorous and provable notion of privacy. Differential Privacy (DP) provides such a framework by ensuring that the influence of any single data point on the algorithm's output is small, typically by adding a carefully calibrated random noise.

19.1.2 Intuition

We compare datasets that differ in exactly one data point. Let D be a dataset and let $D' = D \cup \{x\}$ (or, more generally, any dataset obtained by adding or removing a single record from D). For a randomized training algorithm $\mathscr{A}$, we want the output distributions—i.e., the distributions over model parameters produced when training on D versus D'—to remain close.

A randomized algorithm is said to be ε-differentially private if the addition or removal of one user's data does not significantly alter its output distribution. Informally, the algorithm "behaves almost the same" whether or not any single individual is included in the dataset, with the difference controlled by ε.

19.1.3 Formalism

An algorithm $\mathscr{A}$ is ε-differentially private if, for all neighboring datasets D and D' differing in exactly one entry, and for all measurable subsets S of outputs,

$$\Pr[\mathscr{A}(D) \in S] \leqslant e^{\varepsilon} \Pr[\mathscr{A}(D') \in S].$$

Equivalently, the definition can be expressed as the ratio:

$$\frac{\Pr[\mathscr{A}(D) \in S]}{\Pr[\mathscr{A}(D') \in S]} \leqslant e^{\varepsilon}.$$

Taking natural logarithms (using that ln is monotone increasing), we obtain:

$$\ln\big(\Pr[\mathscr{A}(D) \in S]\big) - \ln\big(\Pr[\mathscr{A}(D') \in S]\big) \leqslant \varepsilon.$$

Thus, ε bounds the *privacy loss*, defined as the log-likelihood ratio between the output distributions of $\mathscr{A}$ on neighboring datasets. A smaller ε implies stronger privacy.

Sensitivity. Let $f : \mathscr{D} \to \mathbb{R}^k$ be a function defined on datasets. The *global sensitivity* of f is defined as

$$\Delta f \;=\; \max_{D,D'} \big\| f(D) - f(D') \big\|_1 ,$$

where the maximum is taken over all **neighboring datasets** D and D', i.e., datasets that differ in exactly one individual's record.

Intuitively, Δf measures the maximum amount by which the value of f can change when the data of a single person is added, removed, or modified.

Example. Consider the function

$$f(D) = \sum_{i=1}^{n} x_i, \qquad x_i \in [0, B].$$

If D and D' are neighboring datasets, then only one value can differ between them, and that value can change by at most B.

Therefore, the global sensitivity is

$$\Delta f = B.$$

19.1.4 Example of a DP Algorithm

Laplace distribution (density). The Laplace distribution with mean μ and scale $b > 0$ has density

$$\mathrm{Lap}(x; \mu, b) = \frac{1}{2b} \exp\Big(- \frac{|x - \mu|}{b} \Big), \qquad x \in \mathbb{R}.$$

Setup. Let $f(D) = \mathrm{COUNT}(D)$ be the counting query. The global sensitivity of f is $\Delta f = 1$, because neighboring datasets D and D' differ in only one record, and hence their counts differ by at most 1.

Define the Laplace mechanism with privacy parameter $\varepsilon > 0$ by

$$\mathscr{A}(D) = f(D) + Z, \qquad Z \sim \mathrm{Lap}(0, b),$$

where we choose the scale parameter $b = \Delta f / \varepsilon = 1/\varepsilon$. Thus the density of the released value $y \in \mathbb{R}$ is

$$p_D(y) = \frac{1}{2b} \exp\Big(- \frac{|y - f(D)|}{b} \Big) = \frac{\varepsilon}{2} \exp\big(-\varepsilon \, |y - f(D)| \big).$$

Theorem. The mechanism $\mathscr{A}(D) = \mathrm{COUNT}(D) + \mathrm{Lap}(1/\varepsilon)$ is ε-differentially private.

Proof. Let D and D' be any neighboring datasets. Since $\Delta f = 1$, we have $|f(D) - f(D')| \leqslant 1$. For any output $y \in \mathbb{R}$, consider the ratio of densities:

$$\frac{p_D(y)}{p_{D'}(y)} = \frac{\frac{\varepsilon}{2} \exp\big(- \varepsilon \, |y - f(D)| \big)}{\frac{\varepsilon}{2} \exp\big(- \varepsilon \, |y - f(D')| \big)} = \exp\Big(\varepsilon \big(|y - f(D')| - |y - f(D)| \big) \Big).$$

Using the inequality $|a| - |b| \leqslant |a - b|$, set $a = y - f(D')$ and $b = y - f(D)$ to obtain

$$|y - f(D')| - |y - f(D)| \;\leqslant\; |f(D) - f(D')| \;\leqslant\; 1.$$

Thus for all y,

$$\frac{p_D(y)}{p_{D'}(y)} \leqslant \exp(\varepsilon).$$

Integrating over any measurable set $S \subseteq \mathbb{R}$ yields

$$\Pr[\mathscr{A}(D) \in S] = \int_S p_D(y)\,dy \leqslant \int_S e^{\varepsilon} p_{D'}(y)\,dy = e^{\varepsilon} \Pr[\mathscr{A}(D') \in S].$$

This is exactly the definition of ε-differential privacy. □

The deterministic mechanism $A_{\text{det}}(D) = \text{COUNT}(D)$ is not differentially private. For neighboring D and D' and the output $y = f(D)$, we have

$$\Pr[A_{\text{det}}(D) = y] = 1 \quad \text{but} \quad \Pr[A_{\text{det}}(D') = y] = 0,$$

so the ratio of probabilities is infinite, violating any finite e^{ε} bound. Adding Laplace noise smooths the output distribution so that the likelihood ratio is always bounded by e^{ε}.

19.2 Composition

Every DP algorithm over a dataset is associated with a privacy cost. Multiple algorithms on same datset can reveal more information by accumulation.

Theorem 19.1 (Basic Composition). *Let $\mathscr{A}_1$ be ε_1-DP and let $\mathscr{A}_2$ be ε_2-DP even when $\mathscr{A}_2$ may be chosen* adaptively *based on the output of $\mathscr{A}_1$. Then the combined mechanism*

$$\mathscr{A}(D) \;=\; \big(\mathscr{A}_1(D),\, \mathscr{A}_2(D,\, \mathscr{A}_1(D))\big)$$

is $(\varepsilon_1 + \varepsilon_2)$-differentially private.

Proof: Fix any pair of neighboring datasets D, D' (differing in one record). Let o_1, o_2 be arbitrary possible outputs of $\mathscr{A}_1, \mathscr{A}_2$ respectively. Write

$$p_D(o_1, o_2) \;=\; \Pr\big[\mathscr{A}_1(D) = o_1,\; \mathscr{A}_2(D, o_1) = o_2\big].$$

By the chain rule for probabilities,

$$p_D(o_1, o_2) \;=\; \Pr[\mathscr{A}_1(D) = o_1] \cdot \Pr\big[\mathscr{A}_2(D, o_1) = o_2 \mid \mathscr{A}_1(D) = o_1\big].$$

Similarly,

$$p_{D'}(o_1, o_2) \;=\; \Pr[\mathscr{A}_1(D') = o_1] \cdot \Pr\big[\mathscr{A}_2(D', o_1) = o_2 \mid \mathscr{A}_1(D') = o_1\big].$$

Because $\mathscr{A}_1$ is ε_1-DP we have for every o_1

$$\Pr[\mathscr{A}_1(D) = o_1] \leqslant e^{\varepsilon_1} \Pr[\mathscr{A}_1(D') = o_1].$$

By assumption $\mathscr{A}_2$ is ε_2-DP even when it is chosen based on the transcript o_1; therefore for every o_1, o_2

$$\Pr\big[\mathscr{A}_2(D, o_1) = o_2 \mid \mathscr{A}_1(D) = o_1\big] \leqslant e^{\varepsilon_2} \Pr\big[\mathscr{A}_2(D', o_1) = o_2 \mid \mathscr{A}_1(D') = o_1\big].$$

Multiplying the two inequalities yields the pointwise bound

$$p_D(o_1, o_2) \leqslant e^{\varepsilon_1} e^{\varepsilon_2}\, p_{D'}(o_1, o_2) = e^{\varepsilon_1 + \varepsilon_2}\, p_{D'}(o_1, o_2).$$

Now let S be any measurable set of output pairs. Integrating (or summing) over $(o_1, o_2) \in S$ gives

$$\Pr[\mathscr{A}(D) \in S] = \int_S p_D(o_1, o_2)\, d(o_1, o_2) \leqslant \int_S e^{\varepsilon_1 + \varepsilon_2}\, p_{D'}(o_1, o_2)\, d(o_1, o_2) = e^{\varepsilon_1 + \varepsilon_2} \Pr[\mathscr{A}(D') \in S].$$

This matches the definition of $(\varepsilon_1 + \varepsilon_2)$-differential privacy.

19.3 Post processing

Theorem 19.2 (Post-Processing). *Let $\mathscr{A}$ be an ε-differentially private mechanism. Let g be any (possibly randomized) function that does* not *access the private dataset. Define the mechanism*

$$\mathscr{B}(D) = g(\mathscr{A}(D)).$$

Then $\mathscr{B}$ is also ε-differentially private.

Proof: Fix neighboring datasets D and D' differing in one individual's data. Let S be any measurable subset of outputs of $\mathscr{B}$. Then

$$\Pr[\mathscr{B}(D) \in S] = \Pr[g(\mathscr{A}(D)) \in S] = \Pr[\mathscr{A}(D) \in g^{-1}(S)],$$

where $g^{-1}(S)$ denotes the preimage of S under g.

Since $\mathscr{A}$ is ε-DP, we have

$$\Pr[\mathscr{A}(D) \in g^{-1}(S)] \leqslant e^{\varepsilon} \Pr[\mathscr{A}(D') \in g^{-1}(S)].$$

Rewriting the term on the right-hand side,

$$\Pr[\mathscr{A}(D') \in g^{-1}(S)] = \Pr[g(\mathscr{A}(D')) \in S] = \Pr[\mathscr{B}(D') \in S].$$

Combining the inequalities gives

$$\Pr[\mathscr{B}(D) \in S] \leqslant e^{\varepsilon} \Pr[\mathscr{B}(D') \in S],$$

which is exactly the definition of ε-differential privacy for $\mathscr{B}$. □

Intuition. Once the data has been passed through an ε-DP mechanism $\mathscr{A}$, its output cannot reveal much about any single individual. Any further computation g, only transforms that already-privacy-protected output. Since g never uses the raw dataset, it cannot "undo" the DP noise or recover additional information about individuals. Hence post-processing cannot worsen privacy beyond ε.

Therefore, after training an LLM on DP algorithm, any subsequent operation such as, alignment, SFT or distillation(applied only to the DP model) is merely a function of the already DP output.

19.4 DP-SGD (Differentially Private Stochastic Gradient Descent)

DP-SGD is the differentially private version of stochastic gradient descent used for training deep learning models while protecting individual training examples. Each iteration applies a DP mechanism to the gradients.

Algorithm (One Iteration)

1. **Sample a minibatch** of training examples.
2. **Compute per-example gradients**
$$g_i = \nabla_\theta \ell(\theta; x_i)$$
for each example x_i in the minibatch.
3. **Clip each gradient** to limit sensitivity:
$$\tilde{g}_i = g_i \cdot \min\left(1,\ \frac{C}{\|g_i\|_2}\right),$$
where $C > 0$ is the clipping threshold. Per-example gradient clipping ensures that one individual's gradient can influence the update by at most a bounded amount (i.e., bounded sensitivity).
4. **Add Gaussian noise** to the averaged clipped gradient:
$$\bar{g} = \frac{1}{B}\left(\sum_{i=1}^{B} \tilde{g}_i + \mathcal{N}\left(0,\ \sigma^2 C^2 I\right)\right).$$
Adding Gaussian noise hides the bounded contribution.
5. **Update model parameters**:
$$\theta \leftarrow \theta - \eta\, \bar{g}.$$

Because every iteration applies a DP mechanism to the gradients, the full training process remains differentially private under the appropriate composition theorem.

19.5 Hypothesis testing

Suppose an observation X is drawn from one of two distributions:

$$X \sim P \quad \text{or} \quad X \sim Q,$$

where P and Q have probability density functions f_P and f_Q. The task is to decide which distribution generated X.

This is a binary hypothesis test:

$$H_0 : X \sim P, \qquad H_1 : X \sim Q.$$

A decision rule is a function that maps the observed value X to either "guess P" or "guess Q".

Error Probabilities

Thw two types of errors that arise are:

- **Type I Error** (False Positive): guessing Q when $X \sim P$.

$$\alpha = \Pr_{X \sim P}(\text{guess } Q).$$

- **Type II Error** (False Negative): guessing P when $X \sim Q$.

$$\beta = \Pr_{X \sim Q}(\text{guess } P).$$

A good test attempts to make both α and β as small as possible. However, reducing one usually increases the other.

Likelihood Ratio and Log-Likelihood Ratio

A fundamental quantity for distinguishing P and Q is the likelihood ratio:

$$\Lambda(X) = \frac{f_P(X)}{f_Q(X)}.$$

Often one uses its logarithm, the *log-likelihood ratio*:

$$T(X) = \log \Lambda(X) = \log\left(\frac{f_P(X)}{f_Q(X)}\right).$$

The value of $T(X)$ indicates whether X is more likely under P or Q.

Neyman–Pearson Lemma

The Neyman–Pearson lemma characterizes the optimal test for deciding between two simple hypotheses P and Q.

Theorem 19.3 (Neyman–Pearson). *Fix a Type I error constraint $\alpha \in [0,1]$. Among all tests with*

$$\Pr_{X \sim P}(\textit{reject } H_0) \leqslant \alpha,$$

the test that minimizes β (Type II error) is the likelihood ratio test, *which has the form:*

$$\textit{Reject } H_0 \textit{ (i.e. guess } Q) \quad \textit{if and only if} \quad T(X) = \log\left(\frac{f_P(X)}{f_Q(X)}\right) < t,$$

for some threshold $t \in \mathbb{R}$ chosen to achieve the desired Type I error level. No other test can achieve a strictly smaller Type II error without increasing the Type I error.

Interpretation

The test compares how likely the observation X is under P versus Q:

$$\Lambda(X) = \frac{f_P(X)}{f_Q(X)}.$$

- If $\Lambda(X)$ is large (or $T(X)$ is large), then X looks more like it came from P.
- If $\Lambda(X)$ is small (or $T(X)$ is small), then X looks more like it came from Q.

Thus the optimal test has the form:

$$\text{Guess } \begin{cases} P, & \text{if } T(X) \geqslant t, \\ Q, & \text{if } T(X) < t. \end{cases}$$

Optimality Insight

The lemma states that:

- Among all tests that keep the probability of rejecting H_0 under P below α, the likelihood ratio test has the smallest possible Type II error β.
- No alternative decision rule can lower both α *and* β simultaneously.

Thus the likelihood ratio test is the unique boundary of optimality for hypothesis testing between two fixed distributions.

Geometric Intuition. Graphically, the optimal decision boundary corresponds to a point where the densities f_P and f_Q intersect when scaled appropriately. Everything to one side is "more P-like" and everything to the other is "more Q-like". This boundary minimizes the area of overlap responsible for both error types.

19.6 Advanced DP

Check [3] for advanced DP composition, tighter accounting than naive ε-summing.

Bibliography

[1] Nicholas Carlini. Extracting training data from large language models. *30th USENIX Security Symposium*, 2021.

[2] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. 2009.

[3] Thomas Steinke. Composition of differential privacy privacy amplification by subsampling. 2022.