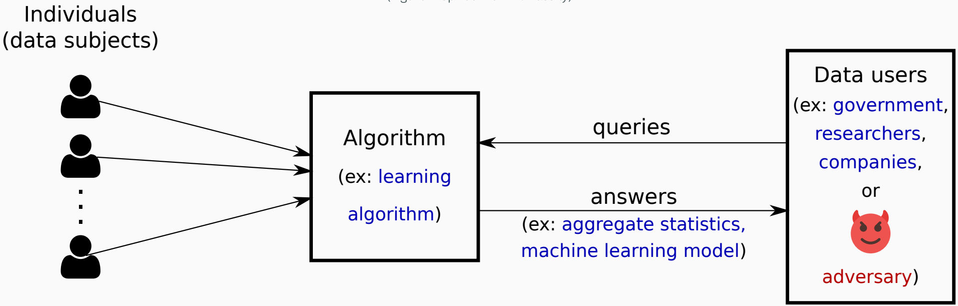
PRIVACY PRESERVING MACHINE LEARNING

LECTURE 4: DIFFERENTIALLY PRIVATE EMPIRICAL RISK MINIMIZATION

Aurélien Bellet (Inria)

Master 2 Data Science, University of Lille

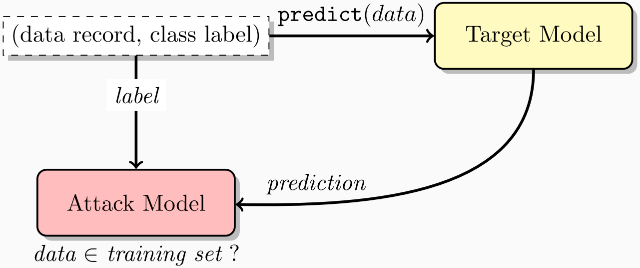
(Figure inspired from R. Bassily)



- We have focused so far on "simple" aggregate statistics
- How about releasing machine learning models trained on private data?

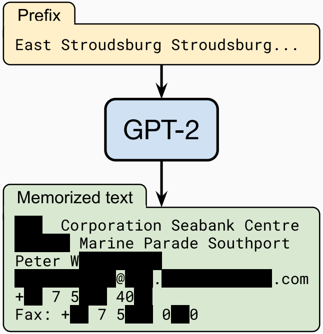
REMINDER: ML MODELS ARE NOT SAFE

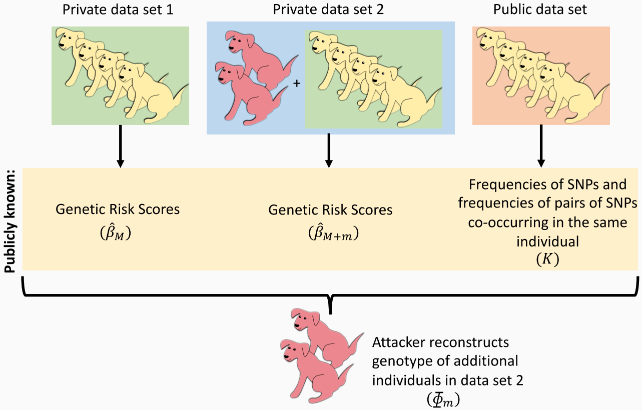
- ML models are elaborate kinds of aggregate statistics!
- As such, they are susceptible to membership inference attacks, i.e. inferring the presence of a known individual in the training set
- For instance, one can exploit the confidence in model predictions [Shokri et al., 2017] [Carlini et al., 2022]



REMINDER: ML MODELS ARE NOT SAFE

- ML models are also susceptible to reconstruction attacks
- For instance, one can extract sensitive text from large language models [Carlini et al., 2021] or run differencing attacks on ML models [Paige et al., 2020]





today's lecture

1. Reminders on Empirical Risk Minimization (ERM)
2. Private ERM via output perturbation

Reminders on Empirical Risk Minimization (ERM)

- For convenience, we focus on supervised learning
- Consider an abstract data space $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the input (feature) space and $\mathcal{Y}$ is the output (label) space
  - For instance, for binary classification with real-valued features: $\mathcal{X} \subset \mathbb{R}^d, \mathcal{Y} = \{-1, 1\}$
- A predictor (model) is a function $h : \mathcal{X} \to \mathcal{Y}$
- We measure the discrepancy between a prediction $h(x)$ and the true label y using a loss function $L(h; x, y)$

- We have access to a training set $D = \{(x_i, y_i)\}_{i=1}^n$ of n data points
- Each data point (x_i, y_i) is assumed to be drawn independently from a fixed but unknown distribution μ
- The goal of ML is to find a predictor h with small expected risk:

$$R(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mu}[L(h; x, y)]$$

- Since μ is unknown, we will use the training set to construct a proxy to R

- We thus define the empirical risk:

$$\hat{R}(h; D) = \frac{1}{n} \sum_{i=1}^{n} L(h; x_i, y_i)$$

- Assume that we work with predictors $h_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^p$
- For notational convenience, we use $L(\theta; x, y)$, $R(\theta)$ and $\hat{R}(\theta)$ to denote $L(h_\theta; x, y)$, $R(h_\theta; D)$ and $\hat{R}(h_\theta; D)$, and omit the dependency on D when it is clear from the context
- Empirical Risk Minimization (ERM) consists in choosing the parameters

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} [F(\theta; D) := \hat{R}(\theta; D) + \lambda \psi(\theta)]$$

- ψ is a regularizer and $\lambda \geq 0$ a trade-off parameter

- We typically work with loss functions that are differentiable in θ: for $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we denote the gradient of L at θ by $\nabla L(\theta; x, y) \in \mathbb{R}^p$
- We also like the loss function, its gradient and/or the regularizer to be Lipschitz

Definition (Lipschitz function)

Let $l > 0$. A function f is l-Lipschitz with respect to some norm $\|\cdot\|$ if if for all $\theta, \theta' \in \Theta$:

$$|f(\theta) - f(\theta')| \leq l\|\theta - \theta'\|.$$

If f is differentiable and $\|\cdot\| = \|\cdot\|_2$, the above property is equivalent to:

$$\|\nabla f(\theta)\|_2 \leq l, \quad \forall \theta \in \Theta.$$

- It is also useful when the loss and/or regularizer are convex or strongly convex

Definition (Strongly convex function)

Let $s \geq 0$. A differentiable function f is s-strongly convex if for all $\theta, \theta' \in \Theta$:

$$f(\theta') \geq f(\theta) + \nabla f(\theta)^\top (\theta - \theta') + \frac{s}{2} \|\theta - \theta'\|_2^2,$$

or equivalently:

$$\left(\nabla f(\theta) - \nabla f(\theta')\right)^\top (\theta - \theta') \geq s \|\theta - \theta'\|_2^2,$$

For $s = 0$, we simply say that f is convex.

EXAMPLE: LOGISTIC REGRESSION

- Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$
- Pick a family of linear models $h_\theta(x) = \text{sign}[\theta^\top x + b]$ for $\theta \in \Theta = \mathbb{R}^p$
- Pick the logistic loss $L(\theta; x, y) = \log(1 + e^{-y(\theta^\top x + b)})$, which is $\|x\|$-Lipschitz and convex
- For $\psi(\theta) = 0$, the ERM problem gives logistic regression
- If we additionally set $\psi(\theta) = \|\theta\|_2^2$, we obtain ℓ_2-regularized logistic regression
- Then $\psi(\theta)$ is 2-strongly convex and $F(\theta) = \hat{R}(\theta) + \lambda\psi(\theta)$ is 2λ-strongly convex

Private ERM via output perturbation

- We would like to privately release a model trained on private data
- A differentially private machine learning algorithm $\mathcal{A} : \mathbb{N}^{|\mathcal{X}\times\mathcal{Y}|} \to \Theta$ should guarantee that for all neighboring datasets D, D' and for all $S_\Theta \subseteq \Theta$:

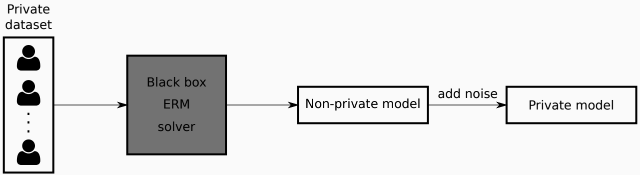
$$\Pr[\mathcal{A}(D) \in S_\Theta] \leq e^\varepsilon \Pr[\mathcal{A}(D') \in S_\Theta] + \delta$$

- **Important note:** in ML, we consider a slightly different neighboring relation where two neighboring datasets $D, D' \in (\mathcal{X} \times \mathcal{Y})^n$ have same size n and differ on one record
  - This corresponds to replacing instead adding/removing one record
  - This is for convenience: normalization term in empirical risk is $1/n$ for both D and D'

- Does DP seem compatible with the objective of ML?
- **Yes!** Intuitively, a model which does not change too much when trained on datasets that differ by a single point should generalize well (because it does not overfit)
- This is related to the notion of algorithmic stability [Bousquet and Elisseeff, 2002], which is known to be a sufficient condition for generalization
- There are formal connections between DP and algorithmic stability [Wang et al., 2016]: in particular, "DP implies stability"

DIFFERENTIALLY PRIVATE ERM VIA OUTPUT PERTUBATION

- ERM is a more complicated kind of "query" than those we have seen so far
- Still, can we re-use some ideas to construct DP-ERM algorithms?
- A natural approach is to rely on output perturbation:



Formally: $\mathcal{A}(D) = \hat{\theta} + \eta$, where $\hat{\theta} \in \arg\min_{\theta \in \Theta}[F(\theta; D) := \hat{R}(\theta; D) + \lambda\psi(\theta)]$

- To calibrate the noise, we need to bound the sensitivity of $\hat{\theta}$
- In some cases, this sensitivity may actually be quite high!
  - Non-regularized objectives with expressive models (e.g., deep neural networks)
  - ℓ_1-regularized models such as LASSO, which are known to be unstable [Xu et al., 2012]

Theorem (ℓ_2 sensitivity for ERM [Chaudhuri et al., 2011])

Let $\Theta = \mathbb{R}^p$. If the regularizer ψ is differentiable and 1-strongly convex, and the loss function $L(\cdot; x, y)$ is convex, differentiable and 1-Lipschitz w.r.t. the ℓ_2 norm for all $x, y \in \mathcal{X} \times \mathcal{Y}$, then the ℓ_2 sensitivity of $\arg\min_\theta F(\theta)$ is at most $2/n\lambda$.

- As expected, sensitivity decreases with n (the size of the dataset)
- Weak regularization leads to large upper bound on sensitivity
- Let's prove this theorem!

Lemma

Let $G(\theta)$ and $g(\theta)$ be two vector-valued functions that are continuous and differentiable everywhere. Assume that $G(\theta)$ and $G(\theta) + g(\theta)$ are λ-strongly convex.

If $\theta_1 = \arg\min_\theta G(\theta)$ and $\theta_2 = \arg\min_\theta G(\theta) + g(\theta)$, then $\|\theta_1 - \theta_2\|_2 \leq \frac{1}{\lambda} \max_\theta \|\nabla g(\theta)\|_2$.

Proof.

- By the optimality of θ_1 and θ_2, we have $\nabla G(\theta_1) = \nabla G(\theta_2) + \nabla g(\theta_2) = 0$
- As $G(\theta)$ is strongly convex, we have $\left(\nabla G(\theta_1) - \nabla G(\theta_2)\right)^\top (\theta_1 - \theta_2) \geq \lambda \|\theta_1 - \theta_2\|_2^2$
- Using Cauchy-Schwartz inequality and the above two results, we obtain:

$$\|\theta_1 - \theta_2\|_2 \|\nabla g(\theta_2)\|_2 \geq (\theta_1 - \theta_2)^\top \nabla g(\theta_2) = \left(\nabla G(\theta_1) - \nabla G(\theta_2)\right)^\top (\theta_1 - \theta_2) \geq \lambda \|\theta_1 - \theta_2\|_2^2$$

- Dividing both sides by $\lambda\|\theta_1 - \theta_2\|$ gives us the result

□

Proof of the theorem.

- Let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $D' = \{(x'_1, y'_1), \ldots, (x_n, y_n)\}$ be two neighboring datasets that differ only in their first point
- Denoting $\hat{\theta} = \arg\min_\theta F(\theta; D)$ and $\hat{\theta}' = \arg\min_\theta F(\theta; D')$, we want to bound $\|\hat{\theta} - \hat{\theta}'\|$
- We define a convenient differentiable function

$$g(\theta) = F(\theta; D') - F(\theta; D) = \frac{1}{n}\Big(L(\theta; x'_1, y'_1) - L(\theta; x_1, y_1)\Big)$$

- By using the 1-Lipschitz property of L we have for any θ:

$$\|\nabla g(\theta)\| = \left\|\frac{1}{n}\Big(\nabla L(\theta; x'_1, y'_1) - \nabla L(\theta; x_1, y_1)\Big)\right\| \leq \frac{2}{n}$$

□

SENSITIVITY BOUND FOR SOME REGULARIZED ERM FORMULATIONS

Proof of the theorem.

- To complete the proof, we will show that $\|\hat{\theta} - \hat{\theta}'\| \leq \frac{1}{\lambda} \max_{\theta} \|\nabla g(\theta)\|$
- Let $G(\theta) = F(\theta; D)$ and recall the definition of $g(\theta) = F(\theta; D') - F(\theta; D)$
- Since L is convex and ψ is 1-strongly convex, $G(\theta)$ and $G(\theta) + g(\theta) = F(\theta; D')$ are λ-strongly convex (as well as differentiable)
- Furthermore, $\hat{\theta}$ and $\hat{\theta}'$ are their corresponding minimizers
- Hence we can apply the lemma, which gives us the desired result

□

Algorithm: DP-ERM via output perturbation $\mathcal{A}_{\text{DP-ERM}}(D, L, \psi, \lambda, \varepsilon, \delta)$

1. Compute ERM solution $\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} F(\theta)$
2. For $j = 1, \ldots, p$: draw $Y_j \sim \mathcal{N}(0, \sigma^2)$ independently for each j, where $\sigma = \frac{2\sqrt{2\ln(1.25/\delta)}}{n\lambda\varepsilon}$
3. Output $\hat{\theta} + Y$, where $Y = (Y_1, \ldots, Y_p) \in \mathbb{R}^p$

Theorem (DP guarantees for DP-ERM via output perturbation)

Let $\varepsilon, \delta > 0$ and $\Theta = \mathbb{R}^p$. Let the loss function L and the regularizer ψ satisfy the conditions of the previous theorem. Then $\mathcal{A}_{\text{DP-ERM}}(\cdot, L, \psi, \varepsilon, \delta)$ is (ε, δ)-DP.

- Proof: a direct application of the Gaussian mechanism with the previous theorem

DP-ERM VIA OUTPUT PERTUBATION: UTILITY GUARANTEES

- Utility is the excess (empirical or expected) risk w.r.t. the non-private solution

Theorem (Utility guarantees for DP-ERM via output perturbation [Chaudhuri et al., 2011])

Consider linear models with $L(\theta; x, y) := L(\theta^\top x, y)$ *and normalized data such that* $\|x\|_2 \leq 1$ *for all* $x \in \mathcal{X}$. *Let* $\psi(\theta) = \frac{1}{2}\|\theta\|_2^2$, $\gamma > 0$ *and* $\beta > 0$. *Let* L *be differentiable and 1-Lipschitz w.r.t. the* ℓ_2 *norm and* ∇L *be 1-Lipschitz w.r.t. the* ℓ_1 *norm. Let* $\theta^* \in \arg\min R(\theta)$ *be a minimizer of the expected risk. If* n *is of order*

$$O\left(\max\left(\frac{\|\theta^*\|_2^2 \log(\frac{1}{\beta})}{\gamma^2}, \frac{p \log(\frac{p}{\beta})\|\theta^*\|_2 \sqrt{\log(\frac{1}{\delta})}}{\gamma\varepsilon}, \frac{p \log(\frac{p}{\beta})\|\theta^*\|_2^2 \sqrt{\log(\frac{1}{\delta})}}{\gamma^{3/2}\varepsilon} \right) \right),$$

then the output θ_{priv} *of* $\mathcal{A}_{\text{DP-ERM}}$ *satisfies* $\Pr[R(\theta_{priv}) \leq R(\theta^*) + \gamma] \geq 1 - 2\beta$.

- The first term in the max is the sample size needed for non-private ERM
- This theorem shows that DP-ERM via output perturbation is well-founded: it matches the utility of the non-private case at the cost of a larger training set

DP-ERM VIA OUTPUT PERTUBATION: DISCUSSION

- An advantage of DP-ERM via output perturbation is that it is simple to implement on top of non-private algorithms
- However it requires restrictive assumptions on the loss function and regularizer
- In practice, ERM is not solved exactly but only to a certain precision using iterative solvers like (stochastic) gradient descent
- Approximate solutions may have small sensitivity, even if no (strongly convex) regularization is used [Zhang et al., 2017]

1. Objective perturbation [Chaudhuri et al., 2011]: output the solution to ERM with a perturbed objective (not covered in the lectures)
2. Gradient perturbation [Bassily et al., 2014, Abadi et al., 2016]: perturb the gradients of a gradient-based algorithm (**next lecture!**)

[Abadi et al., 2016] Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016).
Deep learning with differential privacy.
In *CCS*.

[Bassily et al., 2014] Bassily, R., Smith, A. D., and Thakurta, A. (2014).
Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds.
In *FOCS*.

[Bousquet and Elisseeff, 2002] Bousquet, O. and Elisseeff, A. (2002).
Stability and Generalization.
Journal of Machine Learning Research, 2:499–526.

[Carlini et al., 2022] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. (2022).
Membership inference attacks from first principles.
In *S&P*.

[Carlini et al., 2021] Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. (2021).
Extracting training data from large language models.
In *USENIX Security*.

[Chaudhuri et al., 2011] Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011).
Differentially Private Empirical Risk Minimization.
Journal of Machine Learning Research, 12:1069–1109.

[Paige et al., 2020] Paige, B., Bell, J., Bellet, A., Gascón, A., and Ezer, D. (2020).
Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores.
In *International Conference on Research in Computational Molecular Biology RECOMB.*

[Shokri et al., 2017] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017).
Membership inference attacks against machine learning models.
In *IEEE Symposium on Security and Privacy (S&P).*

[Wang et al., 2016] Wang, Y.-X., Lei, J., and Fienberg, S. E. (2016).
Learning with Differential Privacy: Stability, Learnability and the Sufficiency and Necessity of ERM Principle.
Journal of Machine Learning Research, 17(183):1–40.

[Xu et al., 2012] Xu, H., Caramanis, C., and Mannor, S. (2012).
Sparse Algorithms Are Not Stable: A No-Free-Lunch Theorem.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(1):187–193.

[Zhang et al., 2017] Zhang, J., Zheng, K., Mou, W., and Wang, L. (2017).
Efficient Private ERM for Smooth Objectives.
In *IJCAI*.