Analyzing Privacy Leakage in Machine Learning via Multiple Hypothesis Testing: A Lesson From Fano

Chuan Guo   Alexandre Sablayrolles   Maziar Sanjabi

Meta

December 22, 2023

Recap DP

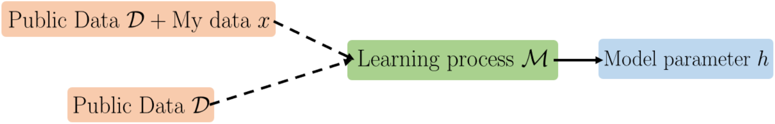


Figure: Make a ambiguity to a learning process $\mathcal{M}$

- A learning process $\mathcal{M}$ report a model h.
- We don't know what data the model is made of.

Recap DP

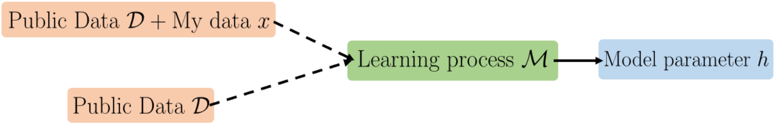


Figure: Make a ambiguity to a learning process $\mathcal{M}$

- A learning process $\mathcal{M}$ report a model h.
- We don't know what data the model is made of.
- Differential privacy is a measurement of an ambiguity on a randomized response.

Recap DP

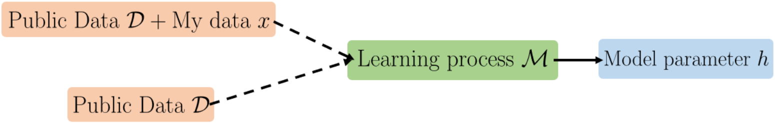


Figure: Make a ambiguity to a learning process $\mathcal{M}$

- A learning process $\mathcal{M}$ report a model h.
- We don't know what data the model is made of.
- Differential privacy is a measurement of an ambiguity on a randomized response.
- ϵ measures uncertainty. For the smaller ϵ, the ambiguity is grower than before.

Recap DP

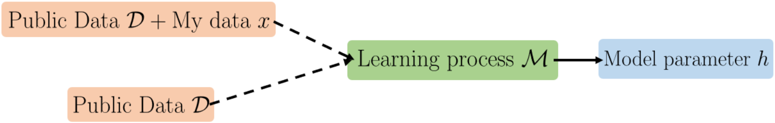


Figure: Make a ambiguity to a learning process $\mathcal{M}$

- A learning process $\mathcal{M}$ report a model h.
- We don't know what data the model is made of.
- Differential privacy is a measurement of an ambiguity on a randomized response.
- ϵ measures uncertainty. For the smaller ϵ, the ambiguity is grower than before.
- People try to create a model where epsilon is small but highly accurate.

Recap DP

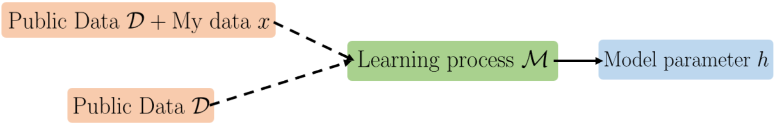


Figure: Make a ambiguity to a learning process $\mathcal{M}$

- A learning process $\mathcal{M}$ report a model h.
- We don't know what data the model is made of.
- Differential privacy is a measurement of an ambiguity on a randomized response.
- ϵ measures uncertainty. For the smaller ϵ, the ambiguity is grower than before.
- People try to create a model where epsilon is small but highly accurate.
- Is there any guidance for a setting ϵ in a limited situation?

Data reconstruction attack

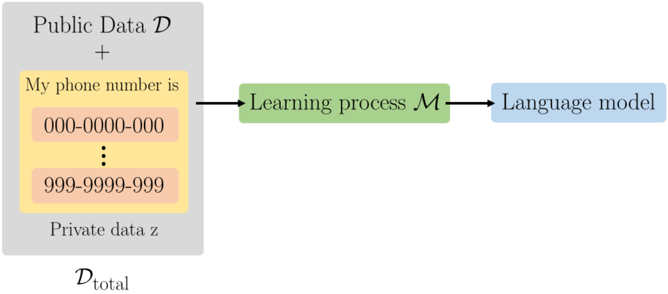


Figure: Learning process of language model with private data

- Let $\mathcal{D}_{\text{total}} = \mathcal{D}_{\text{pub}} \cup \{z\}$ be the training set.

Data reconstruction attack

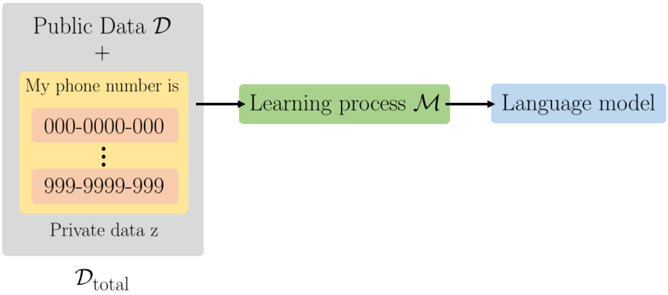


Figure: Learning process of language model with private data

- Let $\mathcal{D}_{\text{total}} = \mathcal{D}_{\text{pub}} \cup \{z\}$ be the training set.
- Private data z is constructed by a nonsensitive data x and a sensitive data u.
- In this case, x is 'My phone number is' and u is a phone number.

Data reconstruction attack

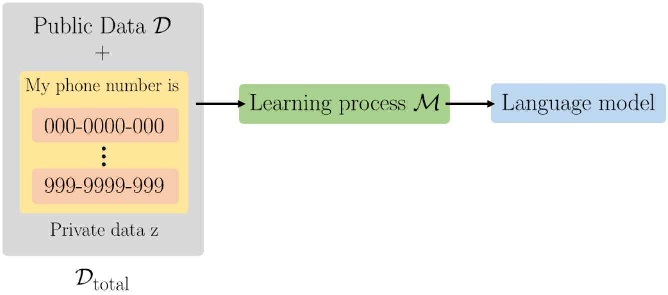


Figure: Learning process of language model with private data

- Let $\mathcal{D}_{\text{total}} = \mathcal{D}_{\text{pub}} \cup \{z\}$ be the training set.
- Private data z is constructed by a nonsensitive data x and a sensitive data u.
- In this case, x is 'My phone number is' and u is a phone number.

Data reconstruction attack

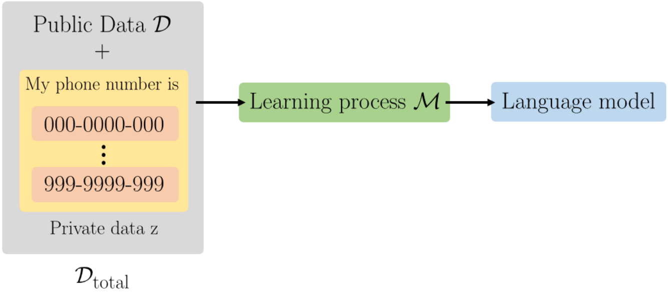


Figure: Learning process of language model with private data

- Let's think from the point of a certain adversary.

Data reconstruction attack

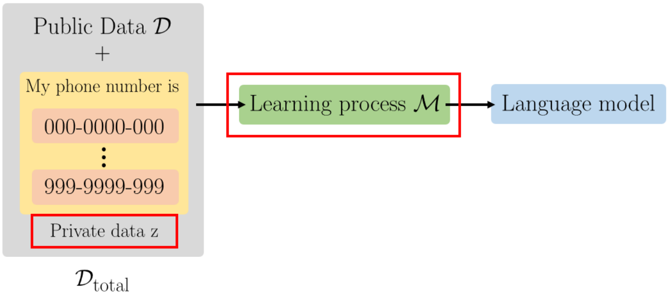


Figure: White box attack scenario

- Let's think from the point of a certain adversary.
- The adversary knows all information except for the ambiguity of the learning process $\mathcal{M}$ and the sensitive data x.

Data reconstruction attack

Figure: Attack scenario for the language model (Calini, 2019)

- For the simple attack scenario, put a 'My phone number is' to a language model
- Then, calculate a likelihood to generate each of the number.

Data reconstruction game

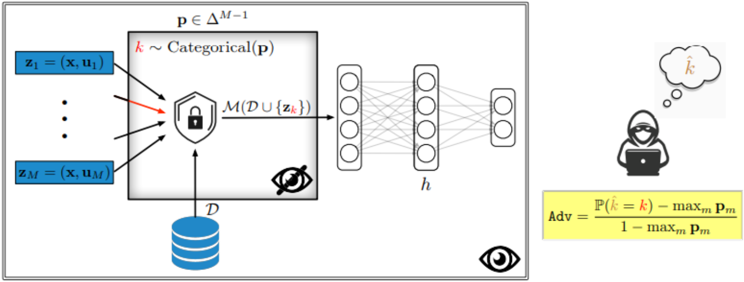


Figure: Illustration of the data reconstruction attack game.

- Let's generalize the situation.

Data reconstruction game

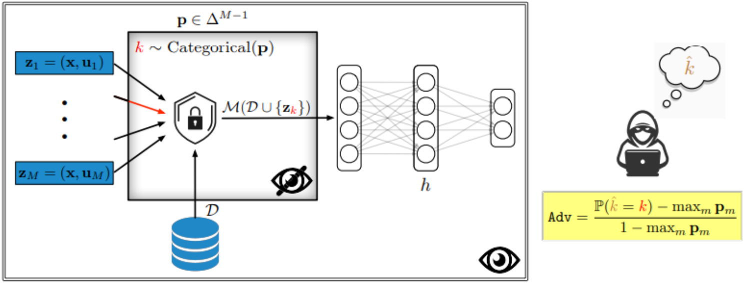


Figure: Illustration of the data reconstruction attack game.

- Let's generalize the situation.
- Recap the previous situation, x is 'My phone number is' and u is a given phone number.

Data reconstruction game

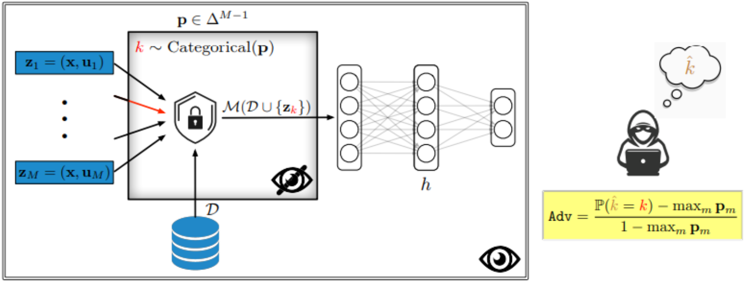


Figure: Illustration of the data reconstruction attack game.

- Let's generalize the situation.
- Recap the previous situation, x is 'My phone number is' and u is a given phone number.
- Make an order the number of candidates 1 to M
- k is a index of the target number.

Data reconstruction game

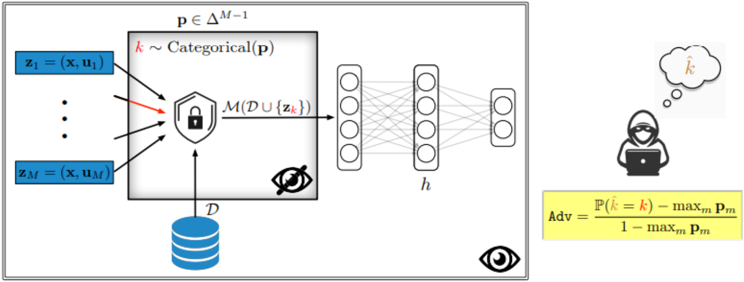


Figure: Illustration of the data reconstruction attack game.

- Adversary wants to know k from the model h.

Advantage metric for the adversary

Next, we define the adversary's goodness metric.

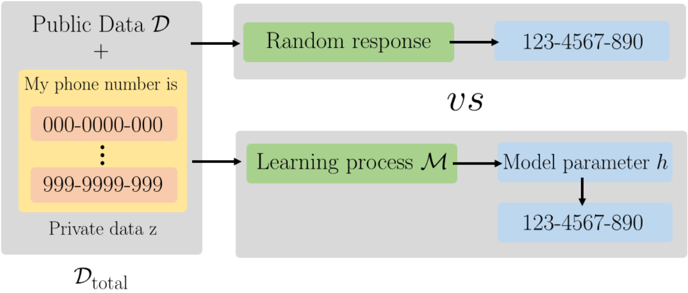


Figure: What security leaks does the model induce?

- In the above situation, when the adversary report $p^* = \max_m p_m$, it is the best scenario.

Advantage metric for the adversary

Next, we define the adversary's goodness metric.

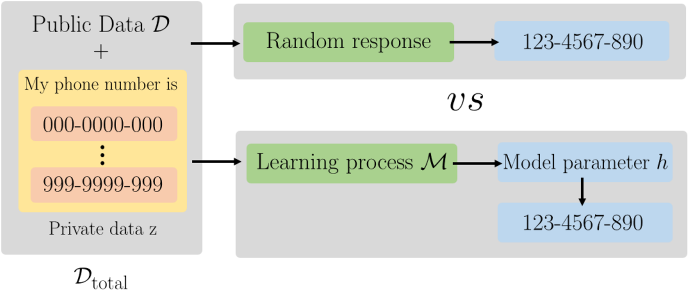


Figure: What security leaks does the model induce?

- In the above situation, when the adversary report $p^* = \max_m p_m$, it is the best scenario.
- p means prior knowledge of the picking number. (ex. the distribution of phone number in Pohang citizens)

Advantage metric for the adversary

Next, we define the adversary's goodness metric.

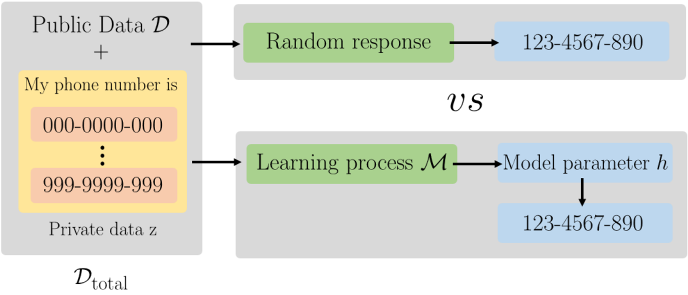


Figure: What security leaks does the model induce?

- For the below scenario, adversary investigates the model h and then infer k.
- We call the probability $P(\hat{k} = k)$ which means the adversary gets the correct number.

Advantage metric for the adversary

$$\mathrm{Adv} = \frac{P(\hat{k} = k) - p^*}{1 - p^*} \in [0, 1] \text{ where } p^* = \max_m \; p_m.$$

- $P(\hat{k} = k)$ is the probability of successfully guessing k upon observing h.
- Advantage metric means information of u, which is additionally exposed by the model.

Fano's inequality

For the Markov chain $k \to \mathcal{M}(\mathcal{D}_{\text{pub}} \cup z_k) \to \hat{k}$, following inequality is satisfied

$$H(k|\mathcal{M}(\mathcal{D}_{\text{pub}} \cup z_k)) \leq H(E) + P(E = 1)log(M - 1),$$

where $E = (\hat{k} \neq k) \in \{0, 1\}$.

- Consequently, we can apply Fano's inequality to bound the adversary's advantage by setting $k = \text{Categorical}(p)$ and differentially private $\mathcal{M}(\mathcal{D}_{\text{pub}} \cup z_k)$.

Fano's inequality

From the Fano's inequality they set

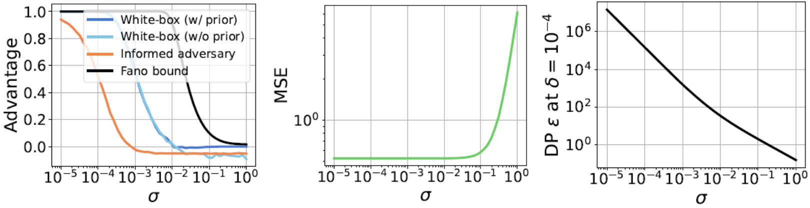
$$f(t) = H(p) - I(k; \mathcal{M}(\mathcal{D}_{\text{pub}} \cup z_k)) \\ + tlogt + (1-t)log(1-t) - tlog(M-1) < 0.$$

where p, M is given and $t = P(E = 1)$.

$$t^* = \min\{t \in [0,1] : f(t) \leq 0\}$$

is calculated to get a upper bound of Adv $\leq (1 - t^* - p^*)/(1 - p^*)$ if we can bound the mutual information term $I(;)$.

DP and mutual information



(a) Advantage vs. σ

(b) Test MSE vs. σ

(c) DP ϵ vs. σ

Figure: Experiments of the data reconstruction attack game.

- The IWPC dataset contains data for clinical trial subjects.
- The goal of an linear regression model to predict the stable dosage of warfarin given the subjects' attributes.
- particularly privacysensitive attribute is the VKORC1 gene type, which can be one of three values: CC, CT or TT.

DP and mutual information

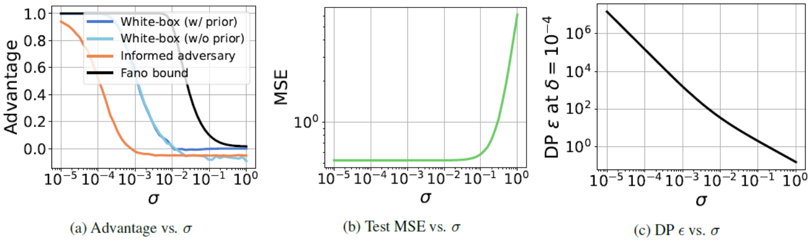


Figure: Experiments of the data reconstruction attack game.

- They use output perturbation on ML model (i.e. $\theta + \mathcal{N}(0, \sigma)$).
- White-box (w/prior) is by MAP.
- White-box (w/o prior) is by MLE.
- Informed adversary is Balle et al. (2022), which only has access to D_{pub} and model θ for predicting the VKORC1 gene type of z.