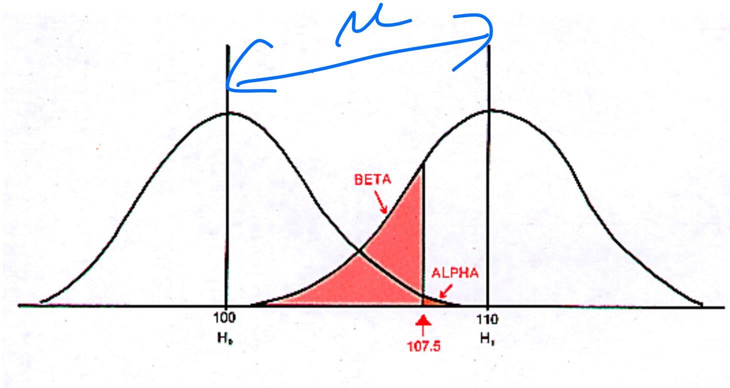
CSCI 699: Privacy Preserving Machine Learning - Week 6

Privacy Auditing and Membership Inference

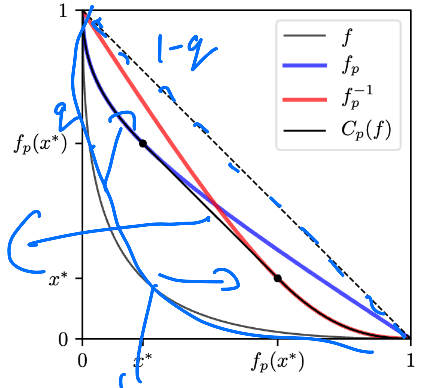
Sai Praneeth Karimireddy, Oct 4 2024

Recap



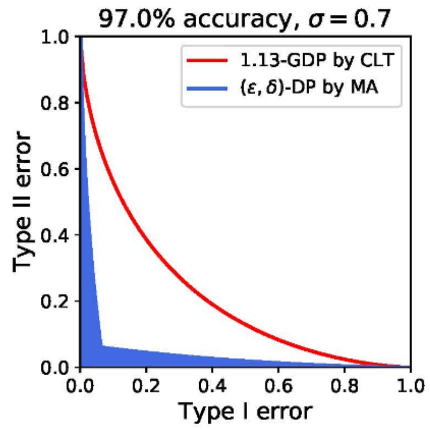
- Gaussian-DP
  - A is μ-GDP if it satisfies f_μ-DP for $f_\mu = T\left(\mathcal{N}(0,1)\ ,\ \mathcal{N}(\mu,1)\right)$
  - Given $f : \mathcal{X}^n \to \mathbb{R}^d$ with Δ bounded ℓ_2-sensitivity, $f(D) + \mathcal{N}\left(0\ ,\ \frac{\Delta^2}{\mu^2} I_d\right)$ is μ-GDP.
  - Composition of $A_1 \circ A_2 \ldots \circ A_k$, each of which is μ_i-GDP is. $\sqrt{\sum_{i=1}^{k} \mu_i^2}$-GDP.

Recap



- Gaussian-DP
  - Composing q-sampling with f-DP, is $\left(\min(f_p, f_p^{-1})\right)^{**}$-DP
  - Central Limit Theorem: $A_1 \circ \ldots \circ A_k$, each satisfying f-DP satisfies μ-GDP $\mu = \frac{2\sqrt{k}\kappa_1}{\kappa_1 - \kappa_2}$ and $\kappa_1 = -\int_0^1 \log|f'(x)|\,dx$ and $\kappa_2 = -\int_0^1 \log^2|f'(x)|\,dx$.
  - K steps of SGD with $\frac{\Delta^2}{\mu^2}I$ variance asymptotically satisfies $\left(q\sqrt{t}\sqrt{e^{\mu^2} - 1}\right)$-GDP $\approx (q\mu\sqrt{t}) - GDP$.

Recap

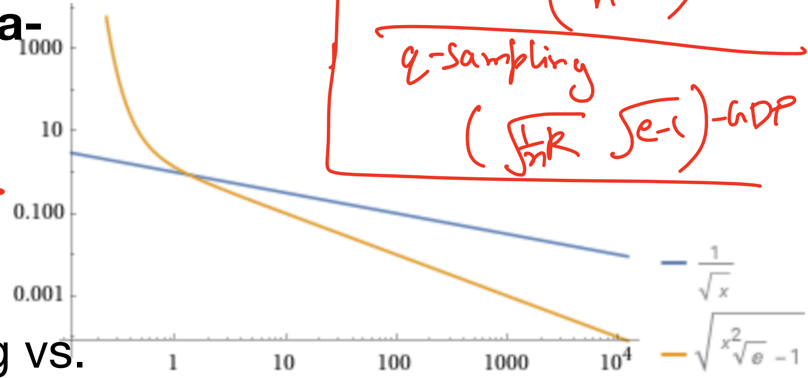


- Subsampled Gaussian mechanism
  - Sample $\mathscr{B}$ where each datapoint is sampled with prob q
  - $\theta_t = \theta_{t-1} - \gamma \left(\left[\frac{1}{|\mathscr{B}|} \sum_{i \in \mathscr{B}} \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_i; \theta), y_i) \right) \right] + \mathcal{N}(0, \tau^2 \rho^2) \right)$
  - After t updates, we have $\left(q\sqrt{t}\sqrt{e^{1/\rho^2} - 1} \right)$-GDP.
  - If k=#epochs, this is $\left(\sqrt{kq}\sqrt{e^{1/\rho^2} - 1} \right)$-GDP.
  - Good default:, set $q = 1/k$ and $\rho = 1$. Gives $\mu = 1.311$.

Recap

Poisson subsampling disadvantages

- $\theta_t = \theta_{t-1} - \gamma \left(\left[\frac{1}{|\mathscr{B}|} \sum_{i \in \mathscr{B}} \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_i; \theta), y_i) \right) \right] + \mathcal{N}(0, \tau^2 \rho^2) \right)$
- I cannot set $\rho \propto |\mathscr{B}|^{-1}$ - mechanism **cannot be data-dependent.**
  It should work for the worst case i.e. when $|\mathscr{B}| = 1$.
- Poisson sampling is a pain - no control over memory.
- Compare for $\rho = 1$. $\left(\sqrt{qk}\sqrt{e-1}\right)$ with subsampling vs.
  k-epochs with full-batch: $\left(\sqrt{k}\sqrt{e^{1/n^2}-1}\right)$.
- But full-batch also does not fit into memory. Use FTRL-DP.

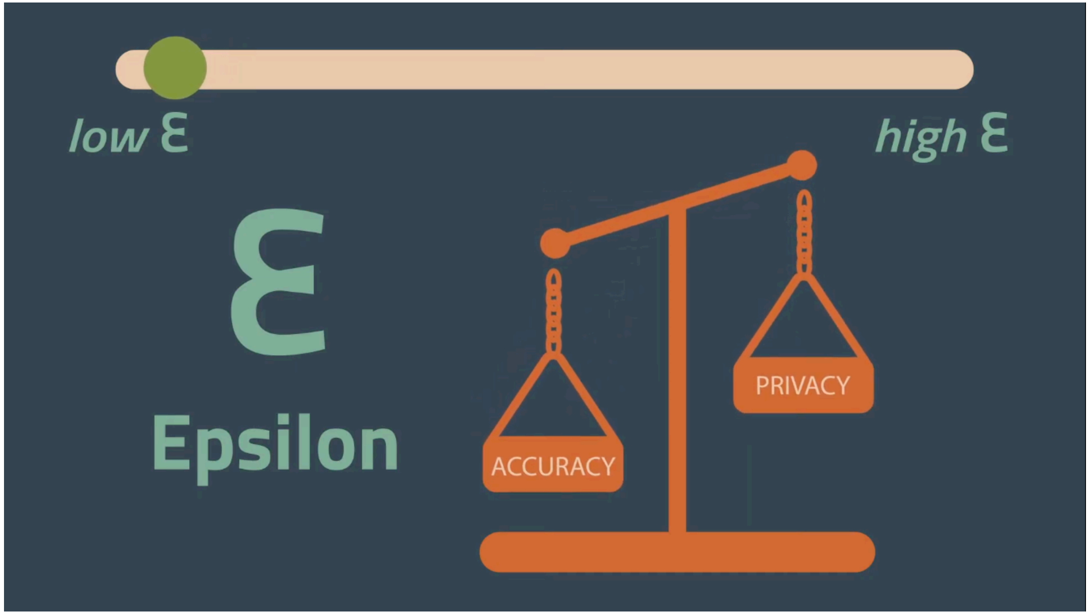


Agenda and announcements

Auditing privacy of ML training

- Privacy Auditing
- Memorization and DP
- Presentations + discussions
- Auditing Practical - HW 3. Postponed to Oct 25 to give you time to focus on projects.
- Remember, Oct 15 deadline for deciding project!
- Next week no class, fall break.

Privacy Auditing Example



Drawbacks of pure theory

- Bounds always loose
    - people assume this and train models with high theoretical ε
- Maybe my implementation is incorrect
- Why should I trust your claim?

Backpropagation Clipping for Deep Learning with Differential Privacy

Timothy Stevens* — *University of Vermont*
Ivoline C. Ngong* — *University of Vermont*
David Darais — *Galois, Inc.*
Calvin Hirsch — *Two Six Technologies*
David Slater — *Two Six Technologies*
Joseph P. Near — *University of Vermont*

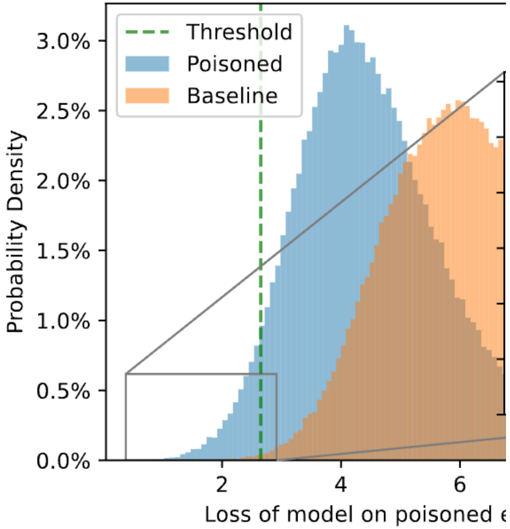
- In 2022, proposed to integrate clipping into forward/backward pass directly
- SOTA accuracy with 30x smaller ε

Privacy Auditing

Debugging Differential Privacy: A Case Study for Privacy Auditing

Florian Tramèr,[*] Andreas Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, Nicholas Carlini
Google Research

- Consider the following test:
  - D = MNIST dataset: 60k images
  - D’ = Add (x', y').
  - Train a CNN θ using [S+22] to get 0.98 acc and (0.21, 10–5)-DP.
  - Check $\ell_\theta(x', y') \leq \tau$. If D’ will be smaller.
  - Repeat 100k on D and 100k on D’.

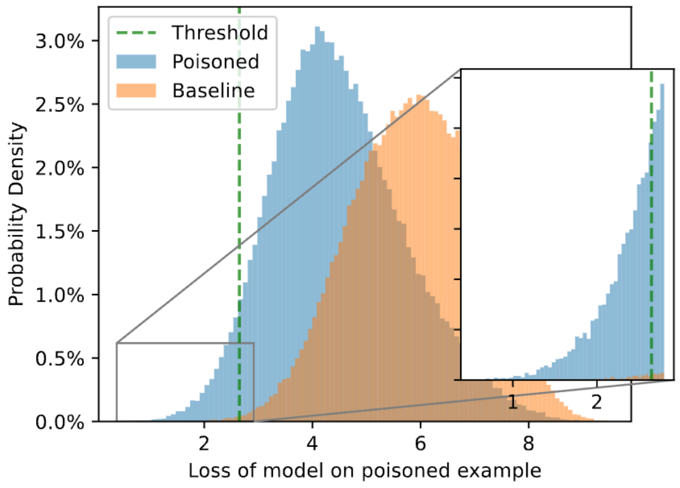


Privacy Auditing

- Some decisions to make
  - Which x'? Called canary
    - insert an *unique* image which model is likely to memorize. Add checkerboard pattern.
  - What y'? Any incorrect label - makes it more unique
  - Try a few images (~25) on an initial 2k training runs.

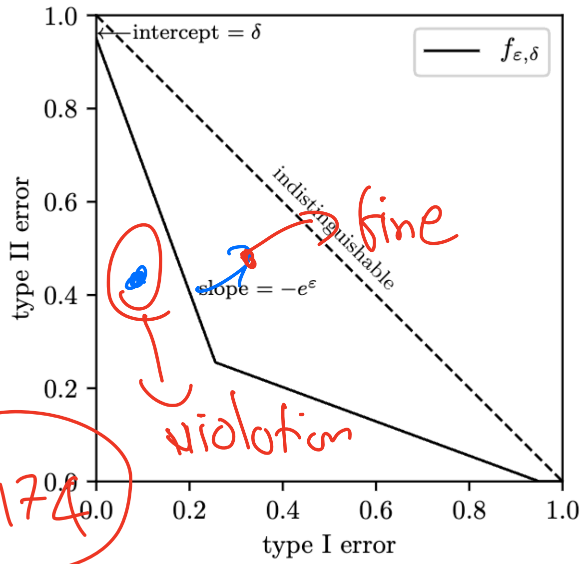
Privacy Auditing

- Some decisions to make
  - Which τ? Can try them all - will get a tradeoff curve.



Privacy Auditing

- Claimed privacy: (0.21, 10–5)-DP.
- With a threshold τ = 2.64 , attack had true positive rate of 4.922% and false positive rate of 0.174%.
- Is this possible?



$$T1E(\alpha) = 0.00174$$

$$T2E(\beta) = 1 - 0.04922$$

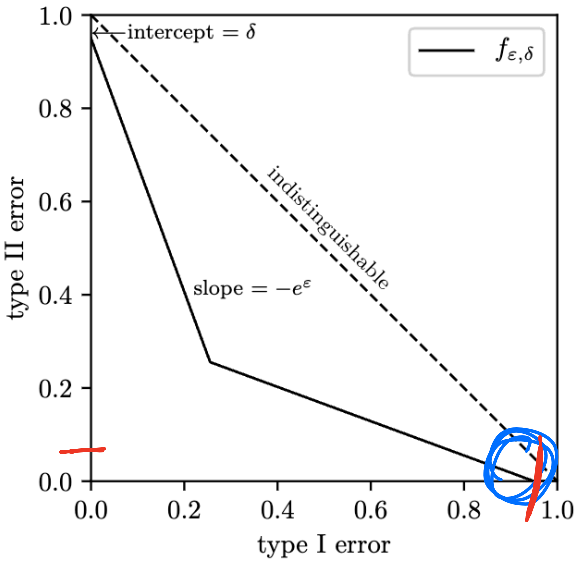
$$\mathbb{E}[\text{guess } H_1 \mid H_0] = \alpha$$

Aside: Clopper-Pearson “exact” method

100K

- $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim$ Bern(α). α is unknown.
- Given Y for n observations, what can we say about α?
- Clopper-Pearson gives intervals $\alpha \in [\alpha^-, \alpha^+]$ with probability $\geq 1 - p$
- No closed form - need to compute numerically.

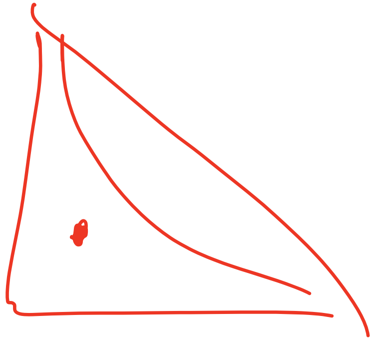
Privacy Auditing



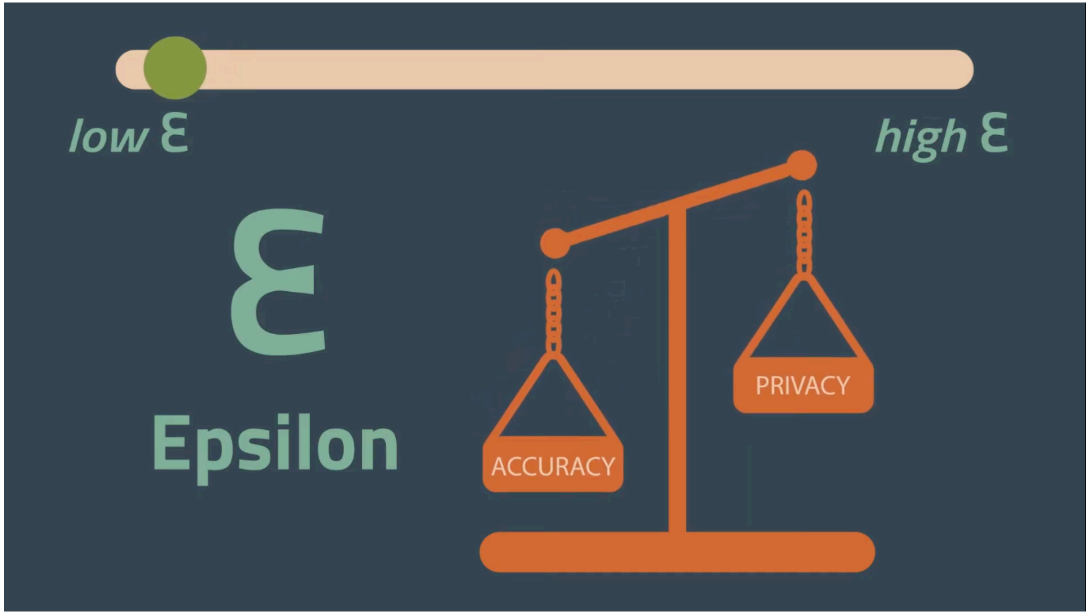
- We have claimed $\hat{\beta} = 0.00174$ and $\hat{\alpha} = 1- 4.922/100 = 0.95078$.
- We have claimed privacy of (0.21, 10–5)-DP.
- $\beta \geq \max(1 - 10^{-5} - e^{0.21}0.95078\ ,\ (1 - 10^{-5} - 0.95078)/(e^{0.21})$ = 0.03988885074

diff is due to sampling?

- Can be due to sampling?
- By Clopper-Pearson, $\alpha^{+} \leq 0.95509$, $\beta^{-} \geq 0.00274$ with $p = 10^{-10}$
- Later, they found a bug and retracted the paper. Very common in DP!!

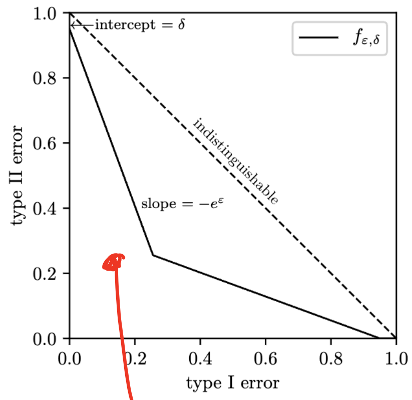


Stronger Audits



Improvements 1: Better Stats

Katz-log intervals



- Do we really need α^+, β^-?
  - We want $\varepsilon = \max\left(\ln(\frac{1-\delta-\beta}{\alpha}), \ \ln(\frac{1-\delta-\alpha}{\beta})\right)$ and α is small.
  - So, we need log of ratio of means of two Bernoulli RVs: $\ln(\frac{1-\delta-\beta}{\alpha})$
  - This turns out to be approximately Gaussian! [Cf. Sec 6.2]. Can get tighter bounds on ε [Lu et al. 23]

$\varepsilon \in [\varepsilon^-, \varepsilon^+]$ with prob atleast 1-p

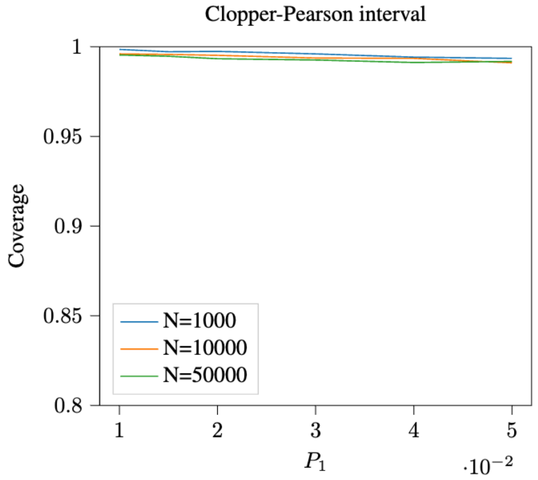
Improvements 1: Better Stats

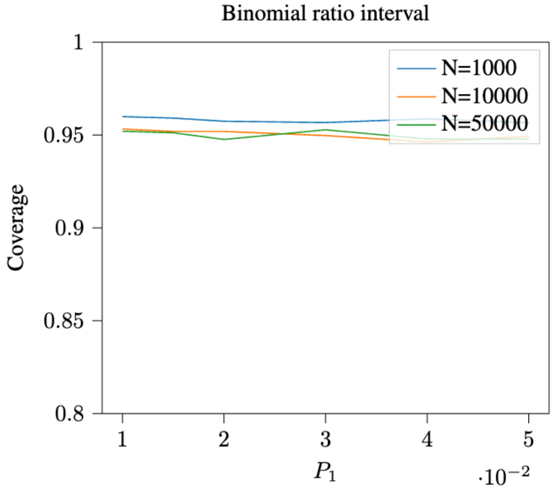
Katz-log intervals

- Consider two Bernoulli RVs with means p_1, p_2, number of trials N and observed values of $n_1$and n_2.

  0.05

  - $$Pr\left(\frac{p_1}{p_2} \notin [\ln\left(\frac{n_1/N}{n_2/N}\right) \pm z_{p/2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}-\frac{2}{N}}]\right) \le p$$
  - where $z_{p/2}$is the critical value of the standard normal (1.96 for α = 0.05).
- Needs to compute ratio of means of two Bernoulli RVs: $\ln(\frac{1-\delta-\beta}{\alpha})$, n_1 = (#false-pos), n_2 = (#true-neg) .

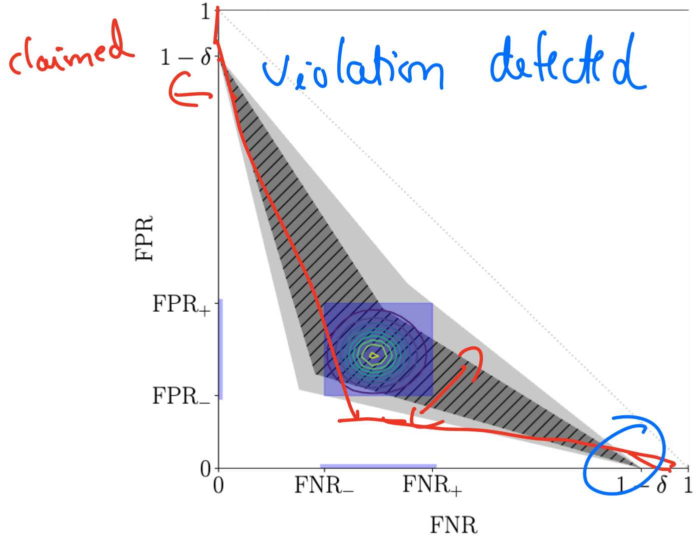


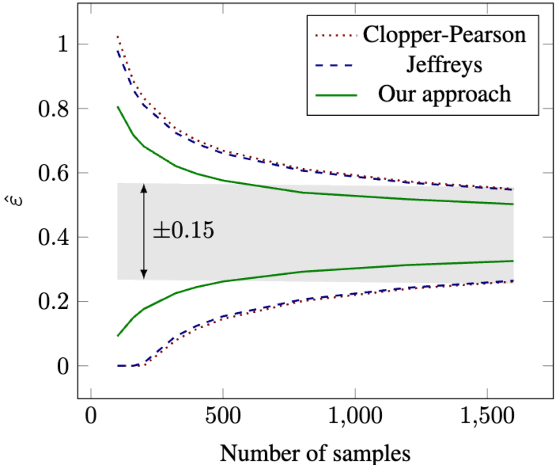


Improvements 1: Better Stats

Bayesian intervals

- Incorporate priors [ZB+23]:
    - Estimate posterior distribution as a Bayesian
    - $\alpha \sim$ Beta$(.5 + n_1, .5 + N - n_1)$, $n_1 =$ #false-pos
      $\beta \sim$ Beta$(.5 + n_2, .5 + N - n_2)$, $n_2 =$ #false-neg
    - Sample lots of (α, β) and compute ε distribution.
    - Reduces number of runs by 3x.
- *Can also use any of your favorite stats tricks.*



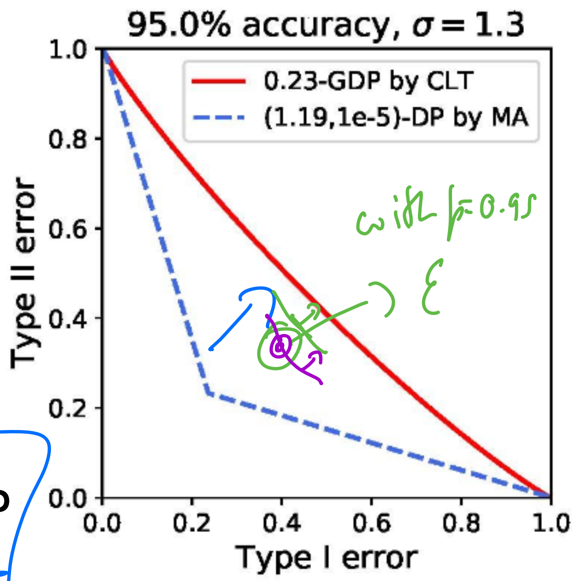


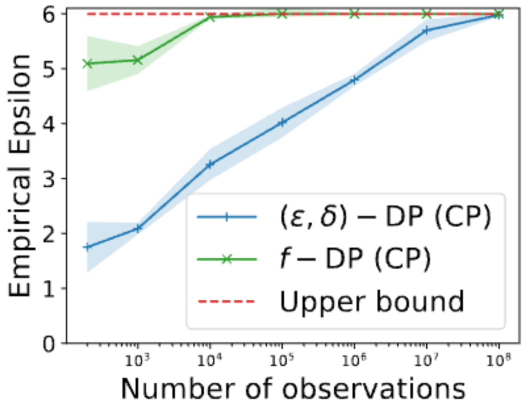
Improvement 2: Use GDP

Gaussian Privacy Auditing

True tradeoff is GDP

- Test for GDP instead:
  - Suppose some Gaussian mechanism claims (ε, δ)-DP
  - Calculate corresponding μ-GDP
  - Check if empirical α, β allows such μ
    $\mu^- = \Phi^{-1}(1-\alpha^+) - \Phi^{-1}(\beta^-)$
  - Reduces number of runs by 10,000x [N+23]



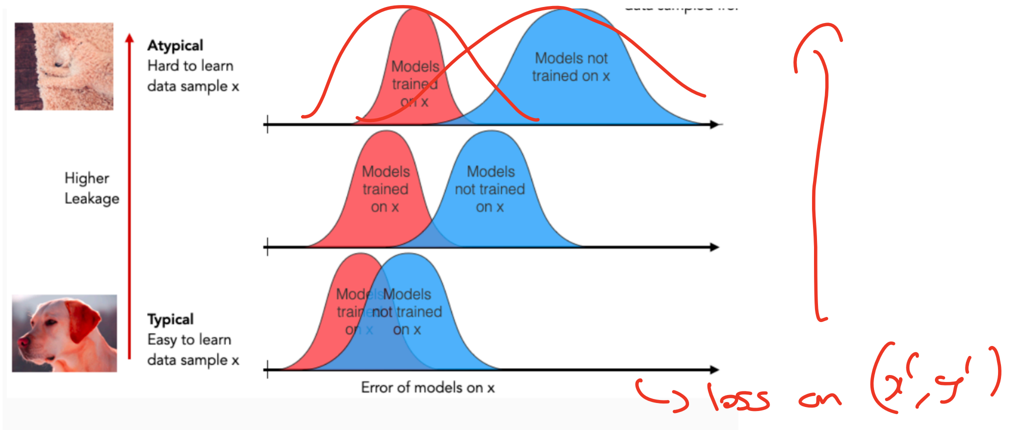


(d) $\varepsilon = 6$

Improvements 3: Better Canaries

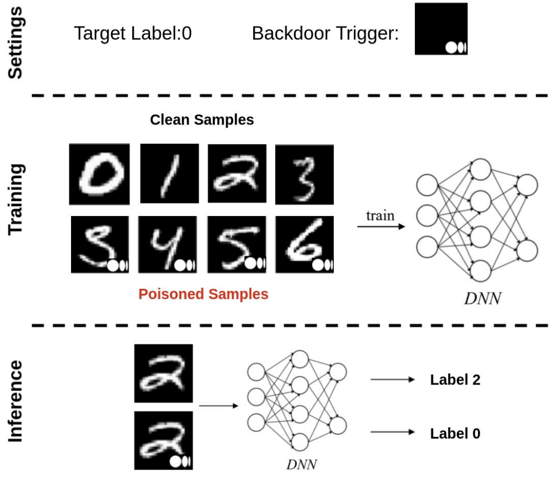
How should you pick the image?

- Picking the right (x', y') is an art
  - Want to add unique/ memorable images

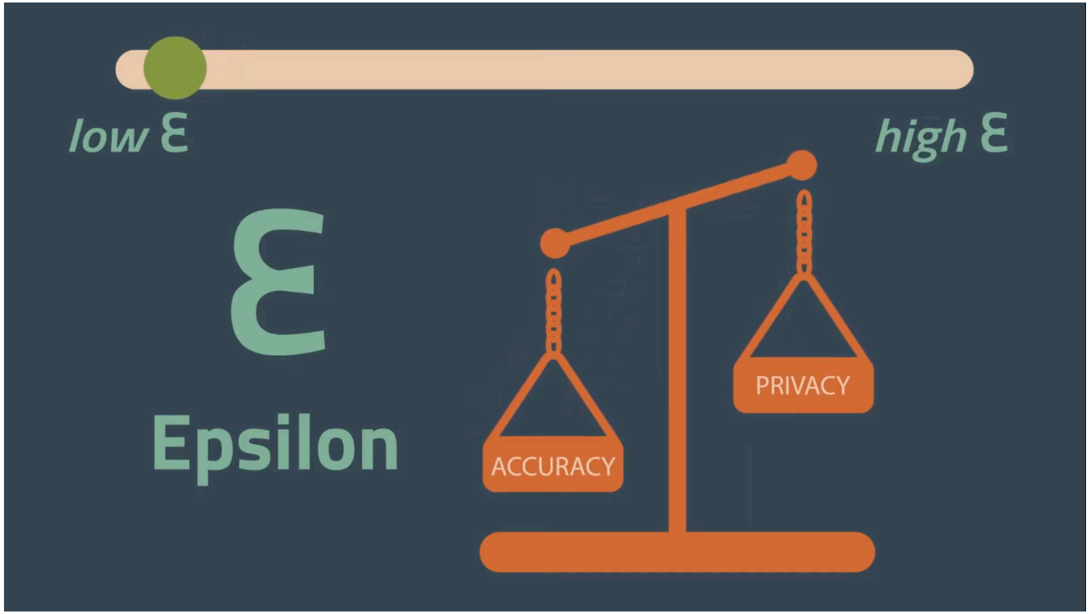


- Insert backdoors / adversarial inputs

$$\max_{\Delta x, \|\Delta x\| \leq \tau'} \|\nabla_\theta \ell(f_\theta(x + \Delta x), y)\|_2$$



Gradient Canaries



Auditing with stronger adversaries

Subsampled Gaussian Mechanism

with q, canary will add a vector of norm C

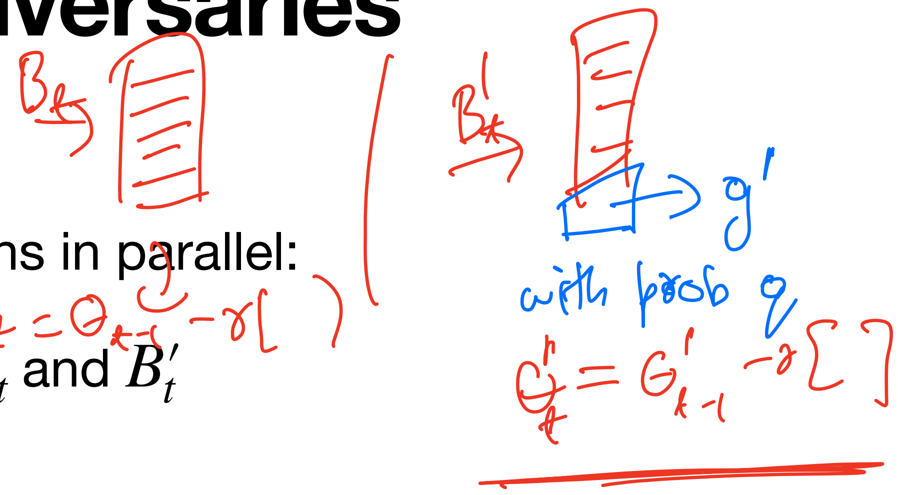
- We know we are running mini-batch gradient descent
  - A mini-batch $\mathcal{B}$ where each datapoint is sampled with prob q
  - Then run,

$$\theta_t = \theta_{t-1} - \gamma \left(\left[\frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_t;\theta), y_i) \right) \right] + \mathcal{N}(0, \rho^2) \right)$$

  - Gradient of canary (x', y') is included with prob. q.
- Mess with gradients directly
  - Instead of editing D, we can directly insert a gradient into update.

Auditing with stronger adversaries

Gradient canary

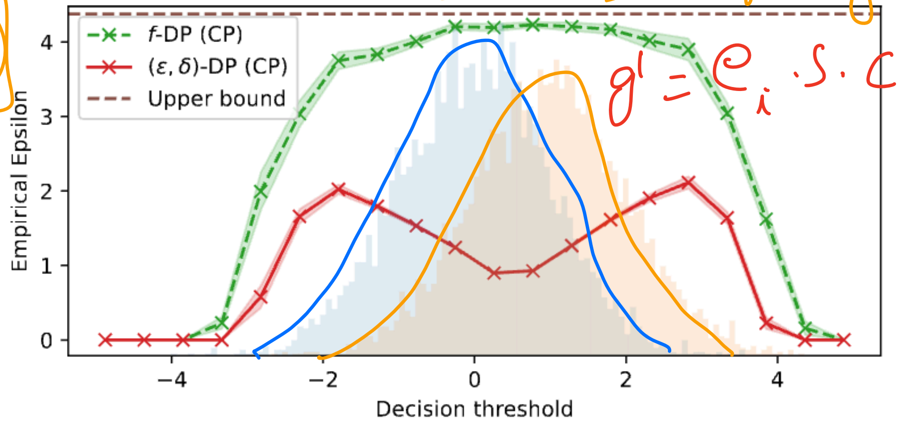


- At each time step t we will run 2 training runs in parallel:
  - Sample 2 batches i.i.d. with prob. q: B_t and B'_t
  - Compute gradients
  - With prob q, add a canary gradient g' to gradients of B'_t
  - Continue private training algorithm
  - Compare $O_t = \nabla_t^\top g'$ and $O'_t = \nabla_t'^\top g'$

Auditing with stronger adversaries

Gradient canary

- Compare $\nabla_t^\top g'$ and $\nabla_t'^\top g'$
- Sample g' randomly - from Gaussian or Dirac
    - In high dimensions, random vectors are orthogonal i.e. we $\nabla_t^\top g' \approx 0$
    - True even after clipping and adding noise
    - But, $\nabla_t'^\top g' \approx \nabla_t^\top g' + q\|g'\|_2 \approx q\tau$
- Gives per-step estimate of ε.
    - Use composition to compute after t-rounds



- Questions: can we
    - simplify to use only a single batch?
    - Use the same g' across t?

Auditing with stronger adversaries

Gradient canary

- Overview [N+23]:
    - Sample g' from Dirac - random coordinate/ Gaussian
    - Estimate posterior distribution of (α, β) using Bayesian method [ZB+23]
    - Estimate per round ε by comparing against sub-sampled Gaussian-DP
    - Combine with composition
- Can detect bugs in noise, clipping, etc. Cannot debug composition.

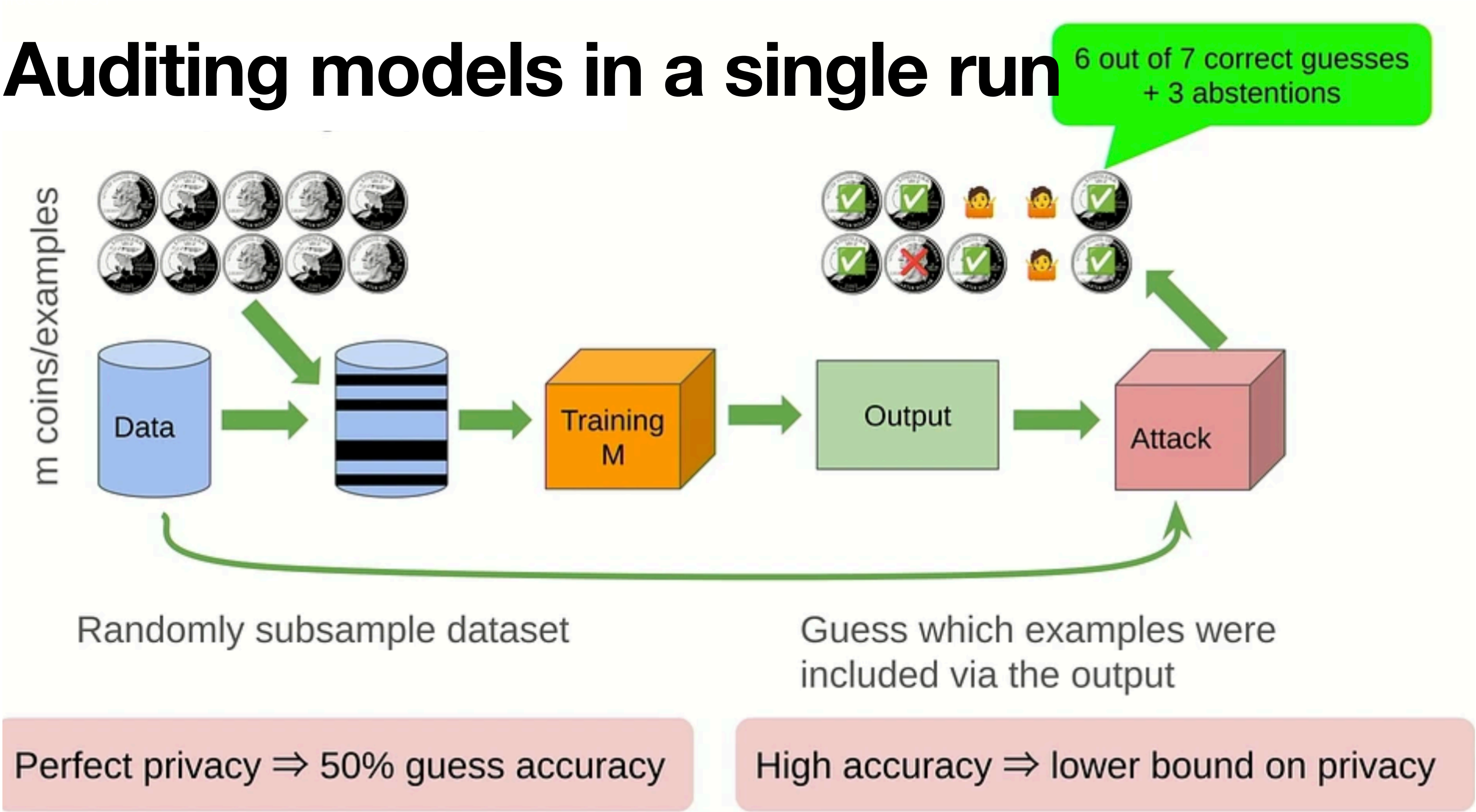
Lower Bounding	Theoretical ε	CIFAR-10 WRN-16
f-DP (CP)	1	0.75
	4	3.40
	8	5.80
	16	11.14
f-DP (ZB)	1	0.95
	4	3.73
	8	7.09
	16	13.95
(ε, δ)-DP (CP)	1	0.41
	4	1.37
	8	3.63
	16	5.25
(ε, δ)-DP (ZB)	1	0.62
	4	2.65
	8	5.07
	16	5.25
ε−DP (Katz)	1	0.49
	4	1.65
	8	4.17
	16	7.52

Auditing models in a single run

Insert multiple canaries

- Gets even better if we insert multiple canaries.
- NeurIPS outstanding paper award! [SNJ23]
- Key idea: insert multiple canary datapoints
  - Include each of m canaries randomly
  - Make m guesses - which canary was present?

Auditing models in a single run



- Overview of auditing scheme [SNJ23]

Auditing models in a single run

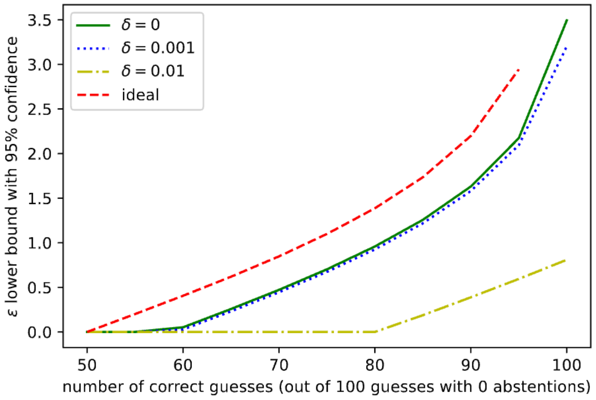
Multiple gradient canaries

- Select a set of canaries: $\mathcal{G} = \{g'_1, \ldots, g'_m\}$.
- For each $i \in [m]$, with prob. 0.5 include $g'_i \in \mathcal{G}'$. Otherwise it is dropped.
- At each time step t:
  - Sample datapoints with prob. q: batch B_t
  - With prob q, add each of the selected canaries g'_i to gradients of B_t
  - Continue private training algorithm
  - Compute: $\{O_i = O_i + \nabla_t^\top g'_i\}$ for $i \in [m]$
- Sort the final $\{O_i\}$, declare top $m/2$ to have been included.

Auditing models in a single run

Multiple gradient canaries

- Relating number of current guesses to ε
- Theorem 5.2 [SNJ23]:
  $Pr[\#\text{ correct guesses} \geq v] \leq Pr\left[\text{Bin}\left(m, \frac{e^{\varepsilon}}{e^{\varepsilon}+1}\right) \geq v\right] + O(\delta)$



Auditing models in a single run

Insert multiple canaries

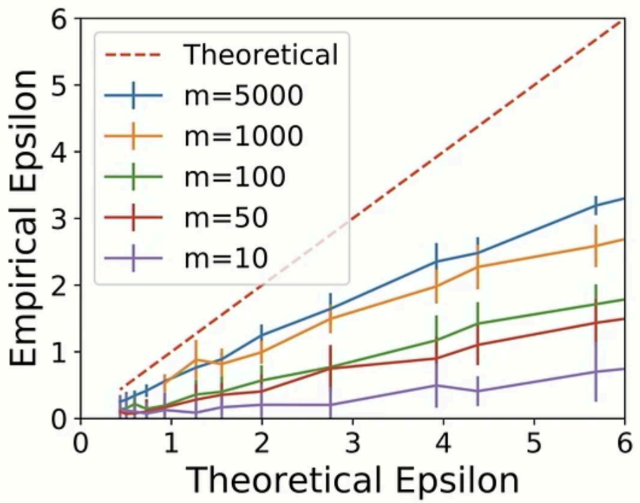


Figure 3. Effect of the number of auditing examples (m) in the white-box setting. By increasing the number of the auditing examples we are able to achieve tighter empirical lower bounds.

Adversary sees intermediate model weights (à la federated learning)

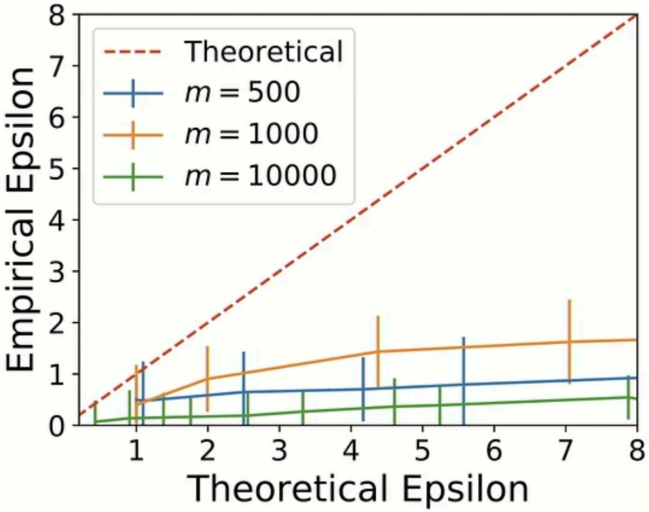
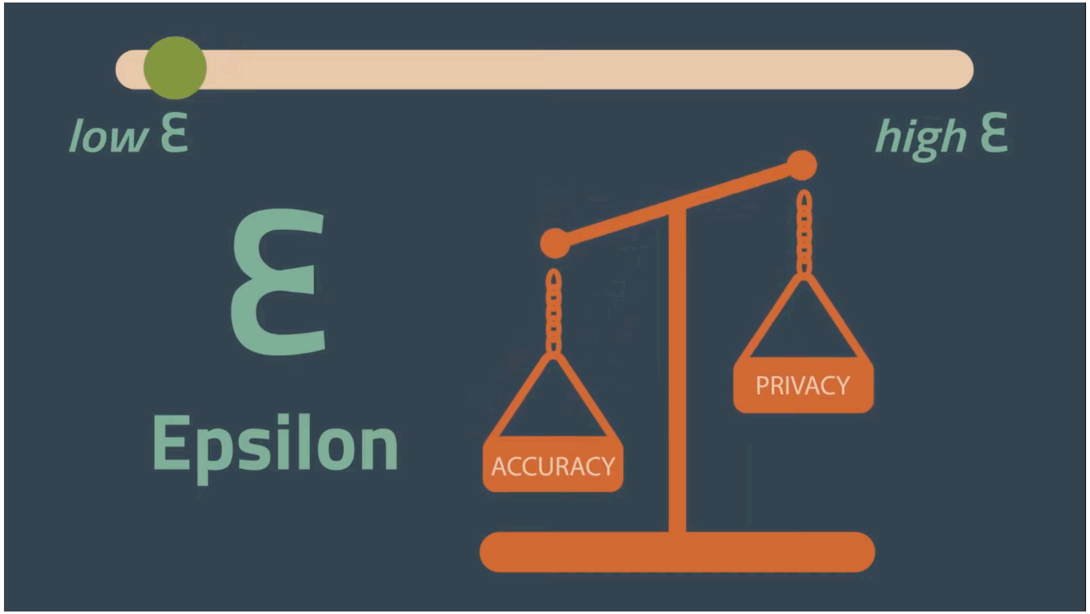


Figure 6. Effect of the number of auditing examples (m) in the black-box setting. Black-box auditing is very sensitive to the number of auditing examples.

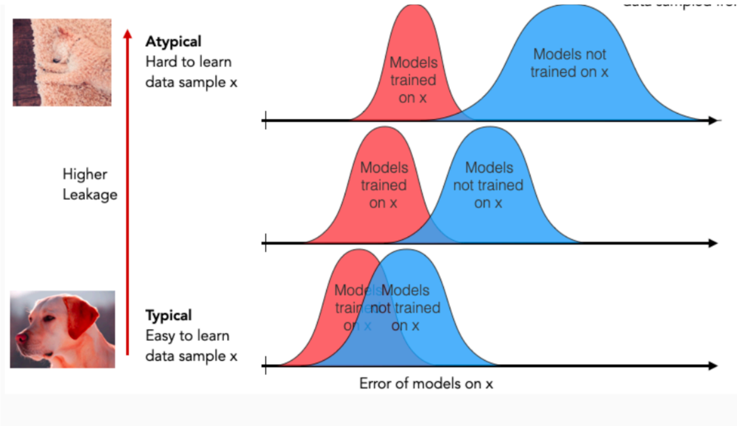
Adversary only sees final model weights (or can only query the loss)

Relaxations of DP



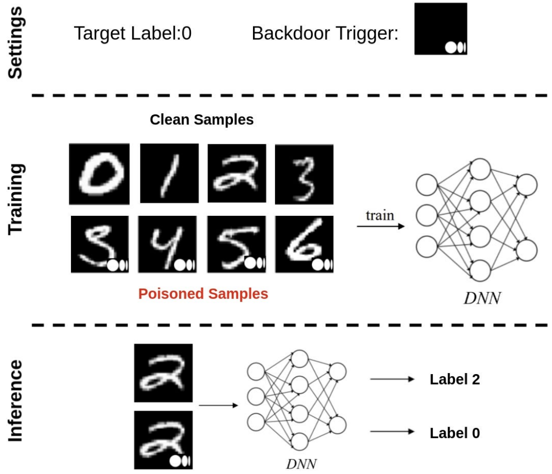
What is a "memorable image"?

- Picking the right (x', y') is an art
  - Want to add unique/ memorable images



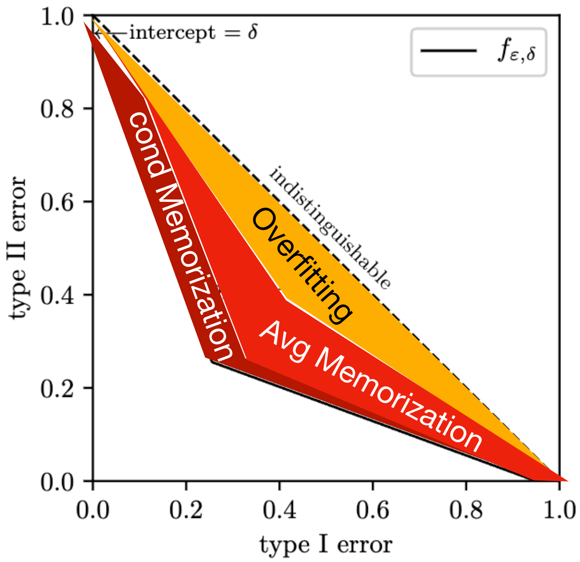
- Insert backdoors / adversarial inputs

$$\max_{\Delta x, \|\Delta x\| \leq \tau'} \|\nabla_\theta \ell(f_\theta(x + \Delta x), y)\|_2$$



Memorization and Privacy

- Overfitting and memorization are both linked to privacy leakage.
- In privacy auditing, we search for memorizing artificial images i.e. search for a “planted signal”. Called *conditional memorization*.
- Avg memorization asks how much of the real training data has been memorized.



Measuring Average Memorization

- Times sued OpenAI claiming they trained on tons of copyrighted data
- For proof, they prompt GPT-4 with the first few paragraphs of an article and then see if it auto-completes an exact match
- 100 instances of match - [exhibit J]

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Case 1:23-cv-11195 Document 1-68 Filed 12/27/23 Page 2 of 127

ONE HUNDRED EXAMPLES OF GPT-4 MEMORIZING CONTENT FROM THE NEW YORK TIMES

Prompt (taken from article):

When Barack Obama joined Silicon Valley's top luminaries for dinner in California last February, each guest was asked to come with a question for the president.

But as Steven P. Jobs of Apple spoke, President Obama interrupted with an inquiry of his own: what would it take to make iPhones in the United States?

Not long ago, Apple boasted that its products were made in America. Today, few are. Almost all of the 70 million iPhones, 30 million iPads and 59 million other products Apple sold last year were manufactured overseas.

Why can't that work come home? Mr. Obama asked.

Mr. Jobs's reply was unambiguous. "Those jobs aren't coming back," he said, according to another dinner guest.

The president's question touched upon a central conviction at Apple. It isn't just that workers are cheaper abroad. Rather, Apple's executives believe the vast scale of overseas factories as well as the flexibility, diligence and industrial skills of foreign workers have so outpaced their American counterparts that "Made in the U.S.A." is no longer a viable option for most Apple products.

Apple has become one of the best-known, most admired and most imitated companies on earth, in part through an unrelenting mastery of global operations. Last year, it earned over $400,000 in profit per employee, more than Goldman Sachs, Exxon Mobil or Google.

However, what has vexed Mr. Obama as well as economists and policy makers is that Apple — and

Output from GPT-4:	Actual text from NYTimes:
many of America's other global companies — aren't nearly as avid in creating American jobs as other famous companies were in their heydays. Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.	many of its high-technology peers — are not nearly as avid in creating American jobs as other famous companies were in their heydays. Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

Defining memorization

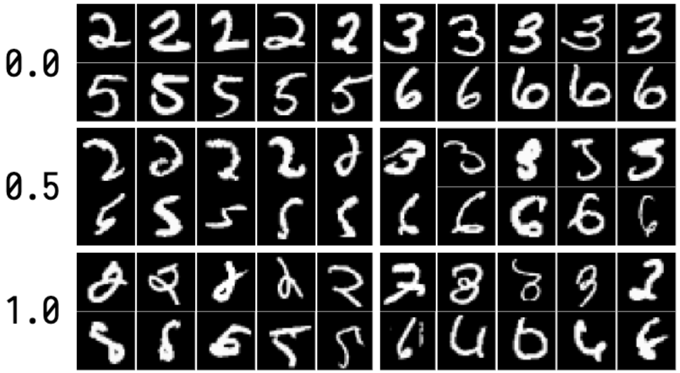
- Memorization: When trained on D, can accurately reconstruct data. If using $D' = D \backslash \{x\}$ cannot. Very useful for weird/tail data.

- **Memorization** [Fel 20] =
$$Pr_{h \leftarrow A(D)}[h(x) = y] - Pr_{h \leftarrow A(D')}[h(x) = y]$$

    - For images: predict labels, in-painting, etc.

    - For text: recover tokens given context

- Memorization ≠ overfitting. k-NN, over-parameterized models memorize exactly. But still generalize.



Most memorized inputs [FZ'20]

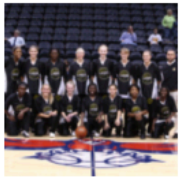
Influence estimation

- **Influence**(x, x_0) =
  $Pr_{h \leftarrow A(D)}[h(x_0) = y_0] - Pr_{h \leftarrow A(D \setminus \{x,y\})}[h(x_0) = y_0]$
  where $h = \arg\min_h E_{x \sim D}[\ell(h(x), y)]$

- Effect of (x, y) on x_0. Many heuristic methods for computing this.

- **Open question**: principled algorithms/ approximation? Proper definitions? Very much understudied.

basketball



basketball

basketball

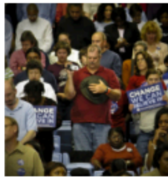
basketball

basketball

volleyball

knee pad

knee pad



cowboy hat

TRAK: [P+'23]

Reference

- How to DP-fy ML: recommended reading for picking a project.

How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy

Natalia Ponomareva *[1], Hussein Hazimeh[1], Alex Kurakin[2], Zheng Xu[2], Carson Denison[3], H. Brendan McMahan[3], Sergei Vassilvitskii[1], Steve Chien[2], and Abhradeep Thakurta[2]

[1]Google Research, NYC
[2]Google Research, MTV
[3]Google Research, Seattle