# Private Ad Modeling with DP-SGD

**AdKDD Workshop, 07 August 2023**

Carson Denison, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Krishna Giri Narra, Amer Sinha, Avinash Varadarajan, Chiyuan Zhang

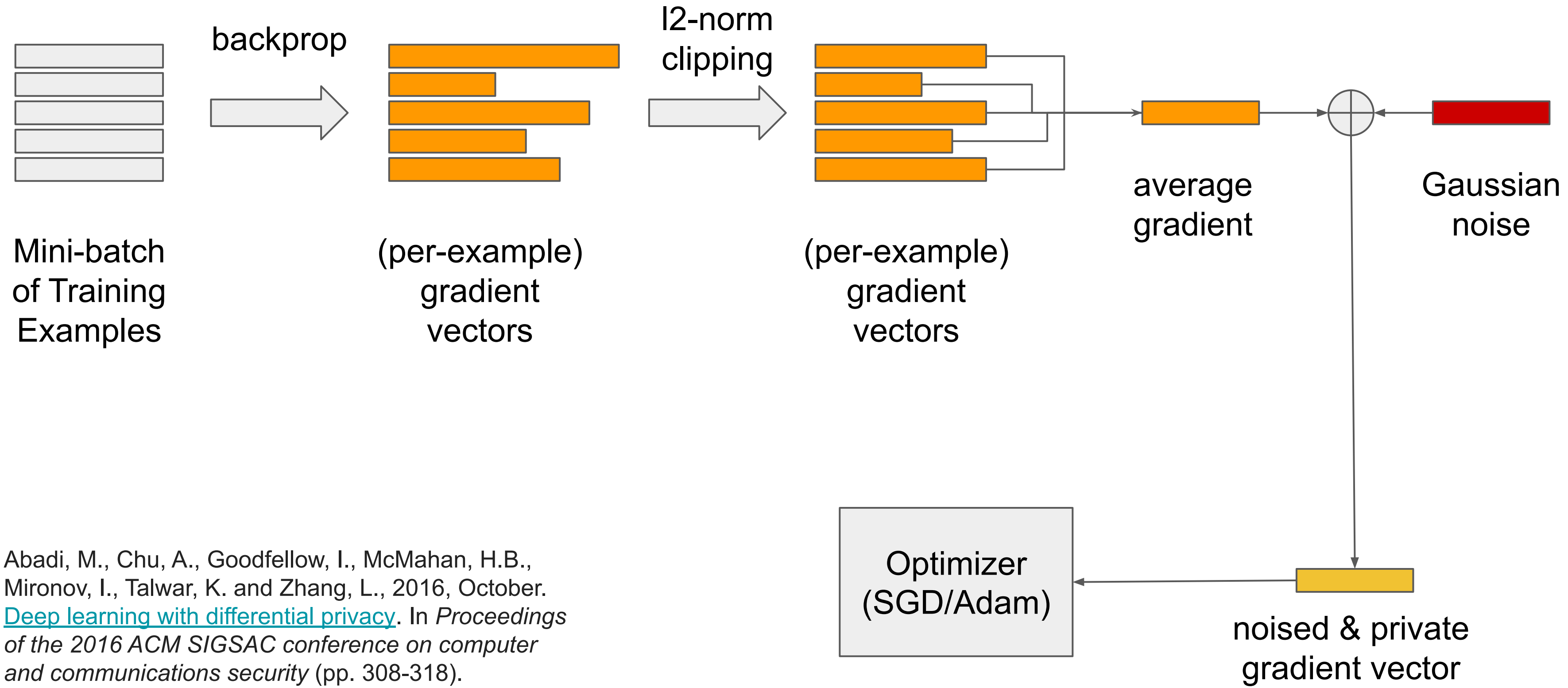


All work done while authors were at Google

# Agenda

# Introduction

# Overview of DP-SGD

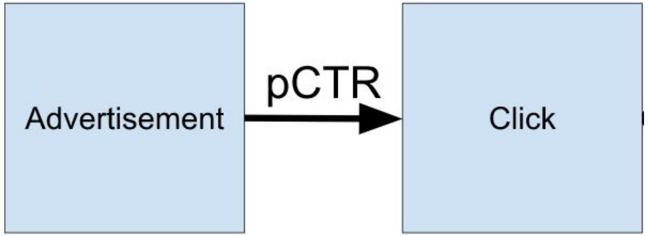



Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L., 2016, October. [Deep learning with differential privacy](). In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 308-318).

# Ads Modeling Overview and Challenges

- Adtechs use models to place ads

- P(Click | Advertisement) - **pCTR**
  - Public Criteo pCTR dataset
  - Binary classification
  - Loss: 1 – AUC (AUC = Area under ROC curve)




- Models are large
  - Billions of parameters
- Data is sparse and class-imbalanced

# What We Contribute

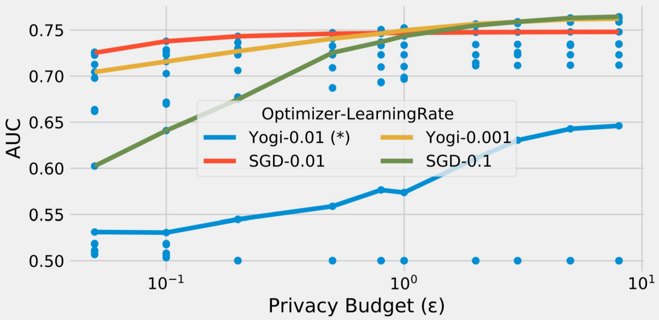
- We show a recipe for training ads models for strong privacy-utility trade off
- We show a simple method for tuning DP-SGD hyperparameters in practice
- We use a new, computationally efficient method for PLD accounting
- We implement DP-SGD that is significantly faster and has low overheads

# Hyperparameter tuning

# Hyperparameter Tuning Overview

- Optimal hyperparameters change!
    - Optimizer
    - Learning Rate
    - Batch size
    - L2 clip norm
- Also depend on privacy budget
    - epsilon (ε) <-> privacy budget

- Batch size and L2 clip norm can be tuned before the others



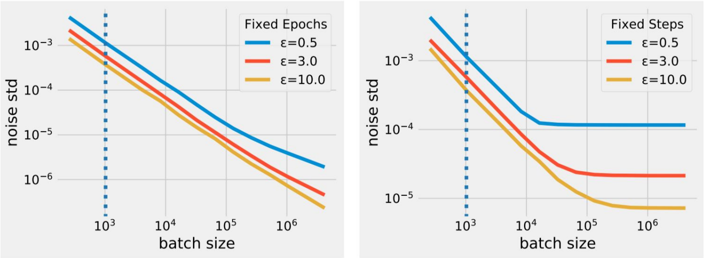

*Best non-private optimizer
Each dot represents the average of 5 runs

# Bigger Batches Need Less Noise

- Noise only added once per batch
  - Bigger batches ⇒ Less noise per example

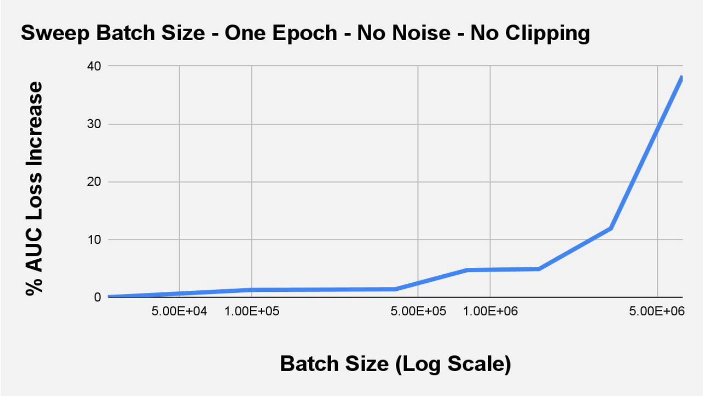
- Large batches often take more epochs to converge

- Can tune batch size before tuning other hyperparameters



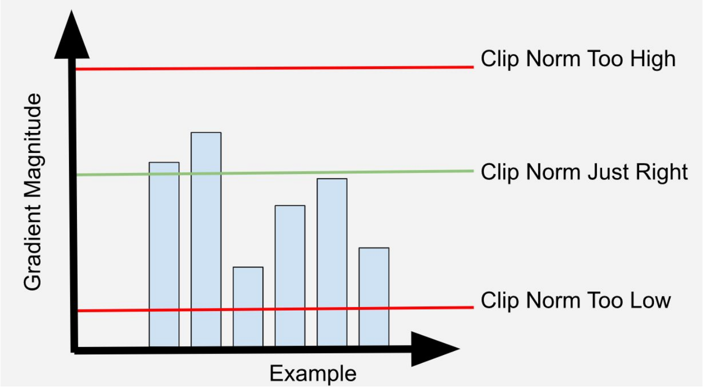

(A) Fixed Epochs (B) Fixed Steps

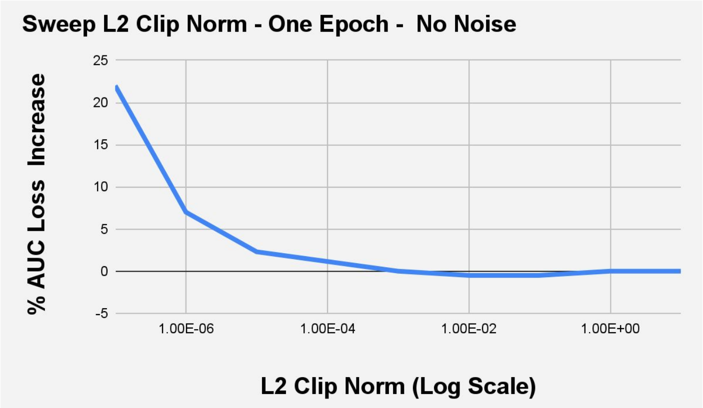
Dotted line shows non-private baseline batch size at various privacy levels



# Clip Norm is a Bias Variance Tradeoff

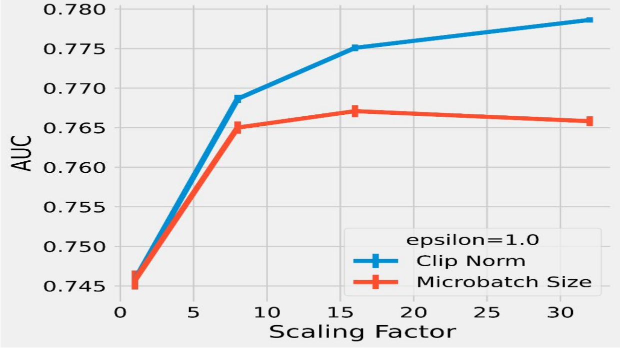
- Noise is scaled with clip norm
- Clipping gradients loses signal
- Tune clip norm using fixed batch size

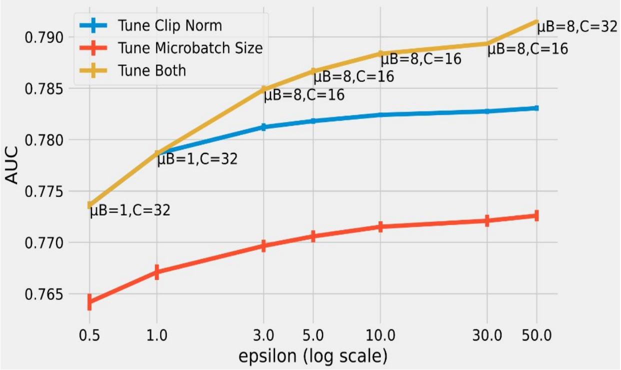


# Micro-batching

- Reduces compute and memory overheads of DP-SGD implementation
- Small microbatch sizes can improve utility
- Clipping and micro-batching help with bias reduction



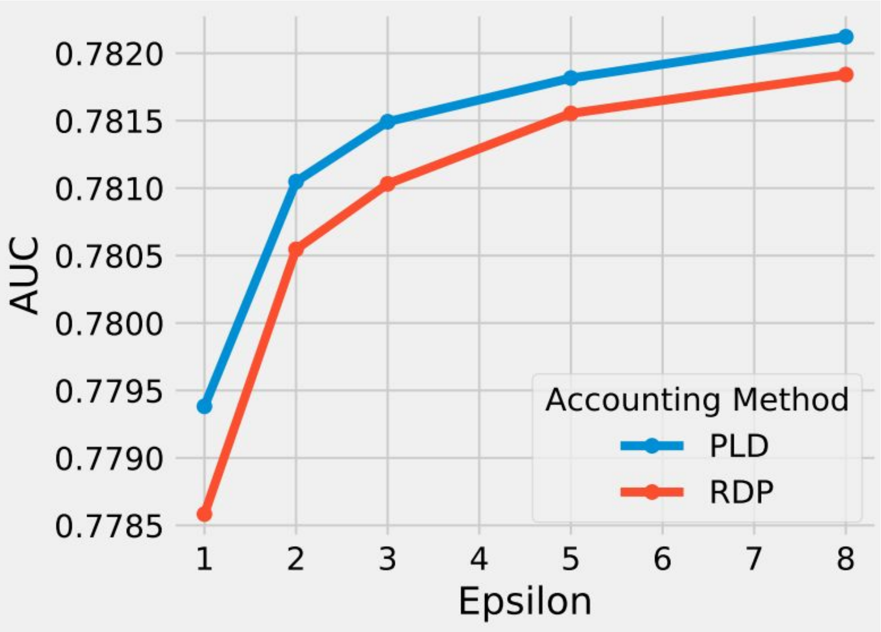
# Tighter Privacy Accounting

# Privacy Loss Distribution (PLD) Accounting

- Privacy loss distribution accounting
  - Tighter than RDP
  - Lots of prior work (see footnotes)[1, 2, 3]
  - Connect-the-dots algorithm is efficient

- https://github.com/google/differential-privacy

- Improves loss by about 0.5%




Loss vs privacy level with **standard Renyi DP** and improved **PLD connect-the-dots** accounting

1. Meiser, S. and Mohammadi, E. Tight on budget? Tight bounds for r-fold approximate differential privacy. In CCS, pp. 247–264, 2018.

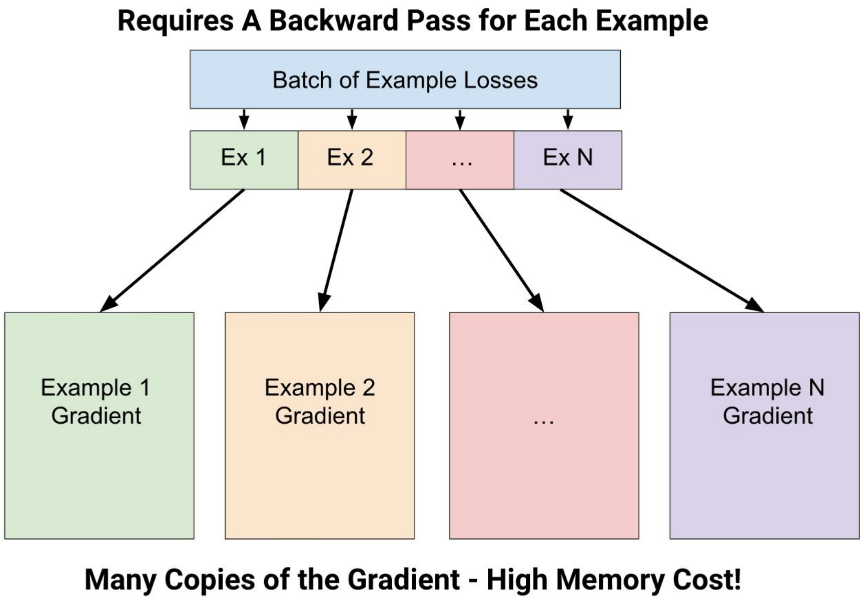
2. Koskela, A., Jalko, J., and Honkela, A. Computing tight differential privacy guarantees using FFT. In AISTATS, pp. 2560–2569, 2020.

3. Doroshenko, V., Ghazi, B., Kamath, P., Kumar, R., and Manurangsi, P. Connect the dots: Tighter discrete approximations of privacy loss distributions. PoPETS, 2022(4): 552–570, 2022.

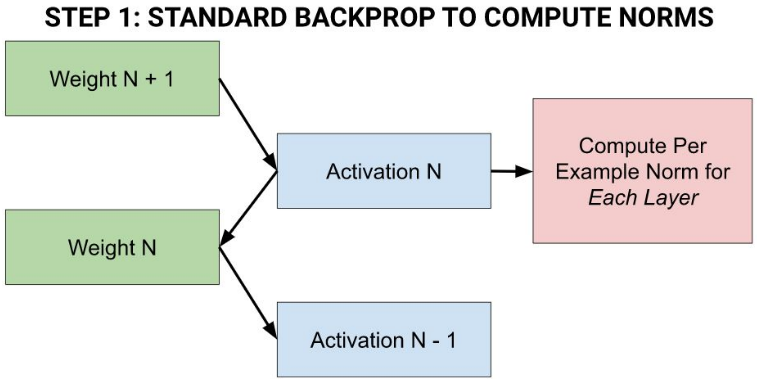
# Efficient Implementation of DP-SGD

# Naive Implementation - Slow and memory inefficient!!!

- Non-private:
  - Max batch size 1,000,000
  - ~**20,000** ex/second
- Naive DP-SGD:
  - Max batch size 50
  - ~**1,000** ex/second

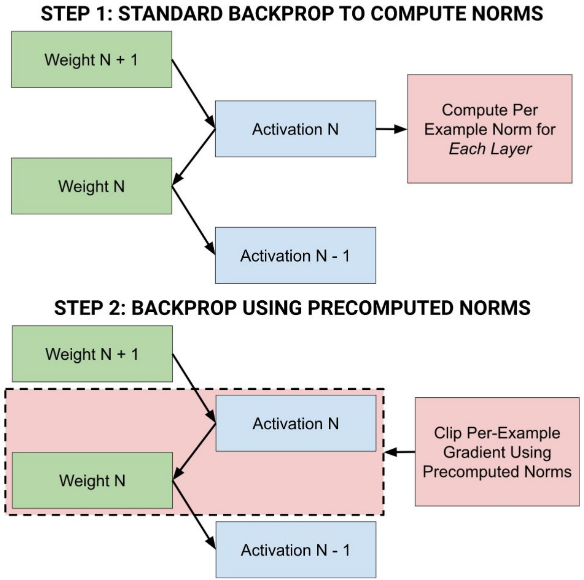


# Careful Implementation of DP-SGD - 20% Slower than Non-Private



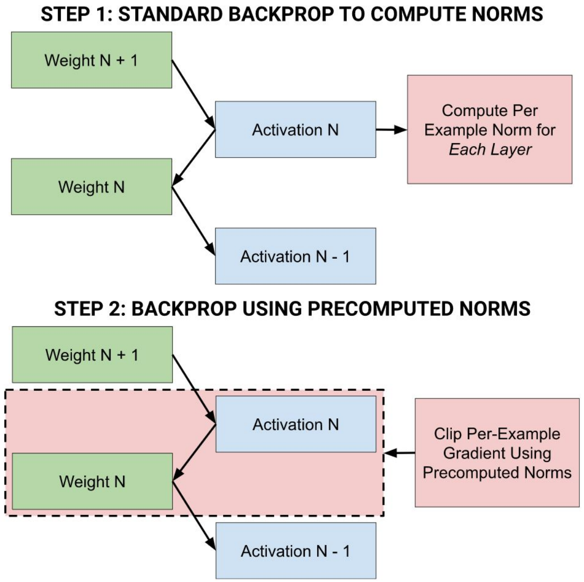

Implementation of gradient norm algorithm from:
Goodfellow, I. Efficient per-example gradient computations.
arXiv preprint arXiv:1510.01799, 2015.

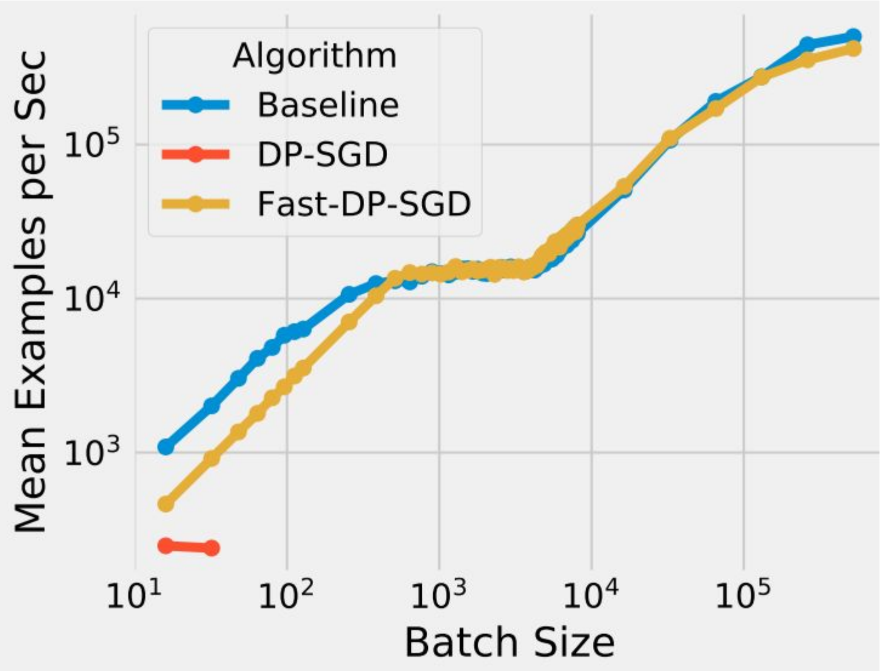
# Careful Implementation of DP-SGD - 20% Slower than Non-Private




Implementation of gradient norm algorithm from:
Goodfellow, I. Efficient per-example gradient computations.
arXiv preprint arXiv:1510.01799, 2015.

# Careful Implementation of DP-SGD - 20% Slower than Non-Private




**Naive DP-SGD** implementation runs out of memory and is orders of magnitude slower than **Fast DP-SGD** or **Non-private Training**

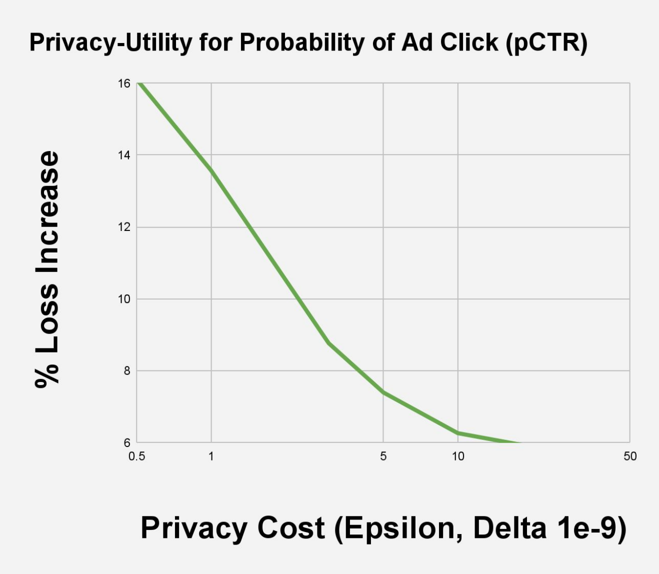
Implementation of gradient norm algorithm from:
Goodfellow, I. Efficient per-example gradient computations. arXiv preprint arXiv:1510.01799, 2015.

# Results

# Results

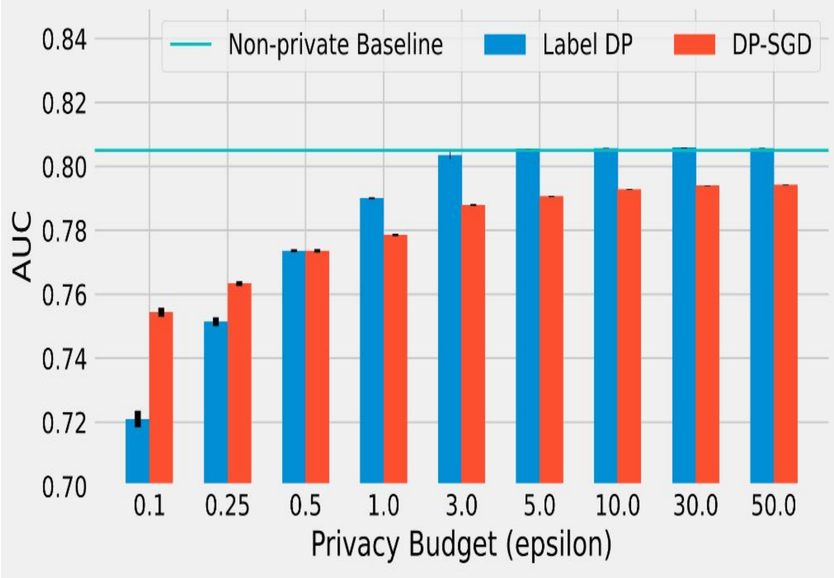
- Competitive Loss with DP-SGD:
  - +6.27% Loss @ epsilon 10
  - +13.58% Loss @ epsilon 1
  - +16.11% Loss @ epsilon 0.5

- Compute needs increased by 20%



# Comparison to Label-DP

- Label-DP
  - Protects privacy of the labels
  - Randomized response mechanism
  - Provides better utility in most regimes
- DP-SGD
  - Protects both inputs and labels
  - Provides better utility in high privacy regimes



# Takeaways

# Takeaways

- Optimal hyperparameters change for private model training
- Carefully implemented DP-SGD is nearly as fast as non-private training
- Competitive privacy-utility trade offs are possible on real-world ads problems

# Q & A

# END