DPIS: An Enhanced Mechanism for Differentially Private SGD with Importance Sampling

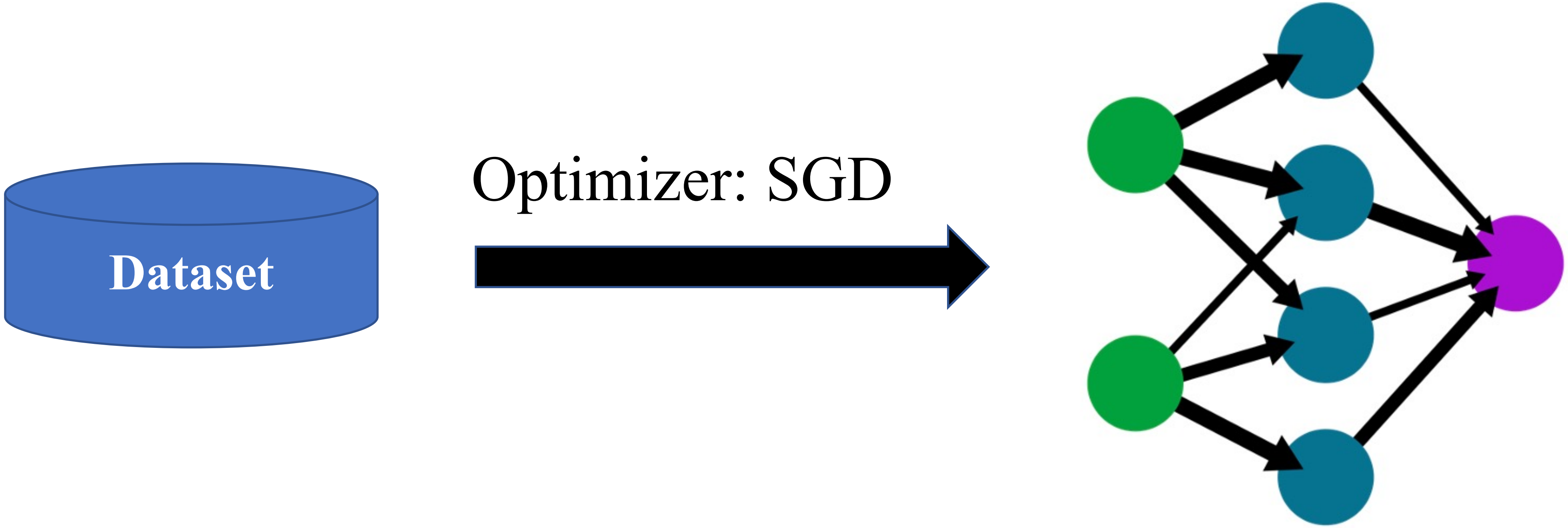
Jianxin Wei, Ergute Bao, Xiaokui Xiao, Yin Yang

NUS National University of Singapore

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY
عضو في مؤسسة قطر
Member of Qatar Foundation

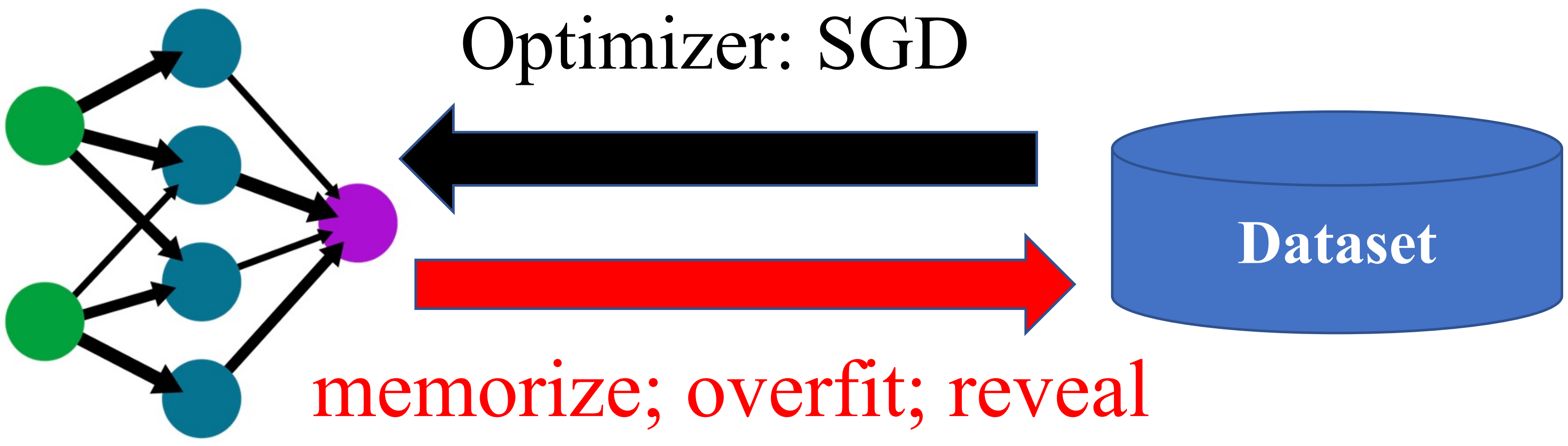
SGD with DP: Motivation

- Goal: train a good neural network on an input dataset using SGD.



SGD with DP: Motivation

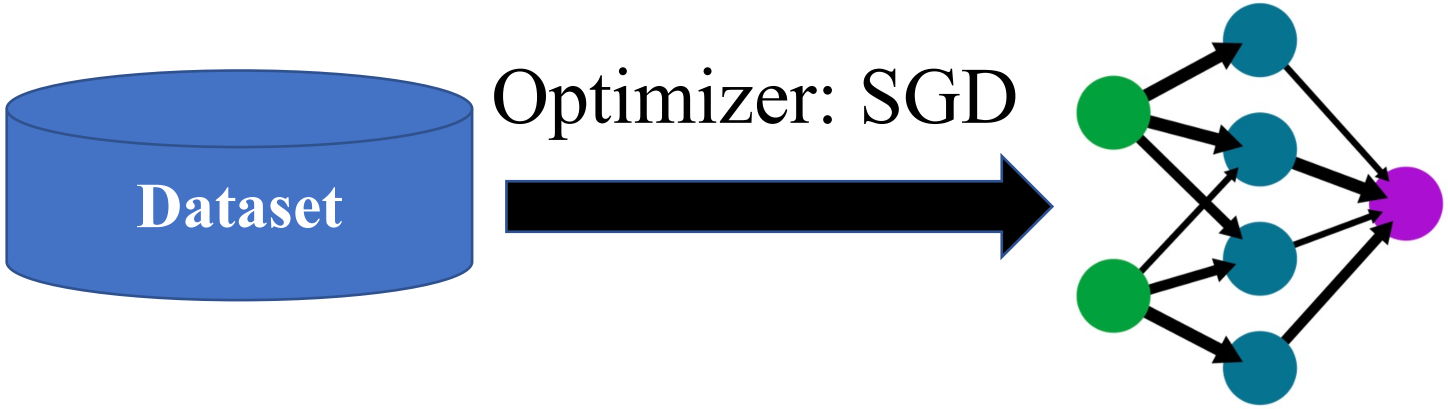
- The trained model is released as a white box.

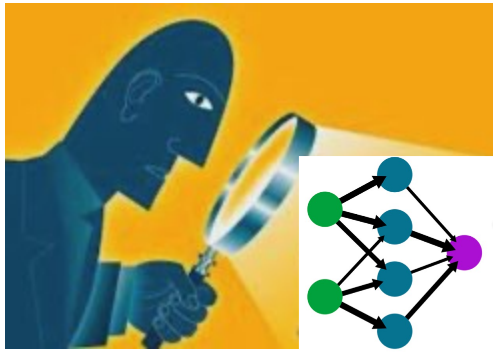


[Shokri et al. S&P' 17] [Carlini et al, USENIX' 18]…

SGD with DP: Problem Setting

- Utility goal: train a good model.
- Privacy goal: protect the data.

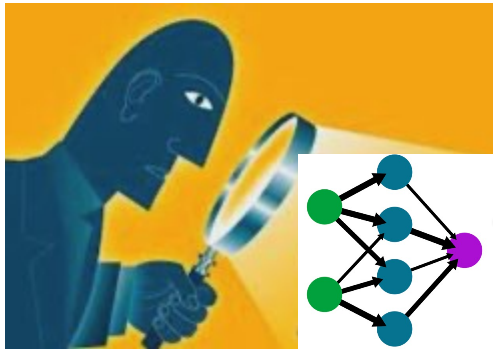


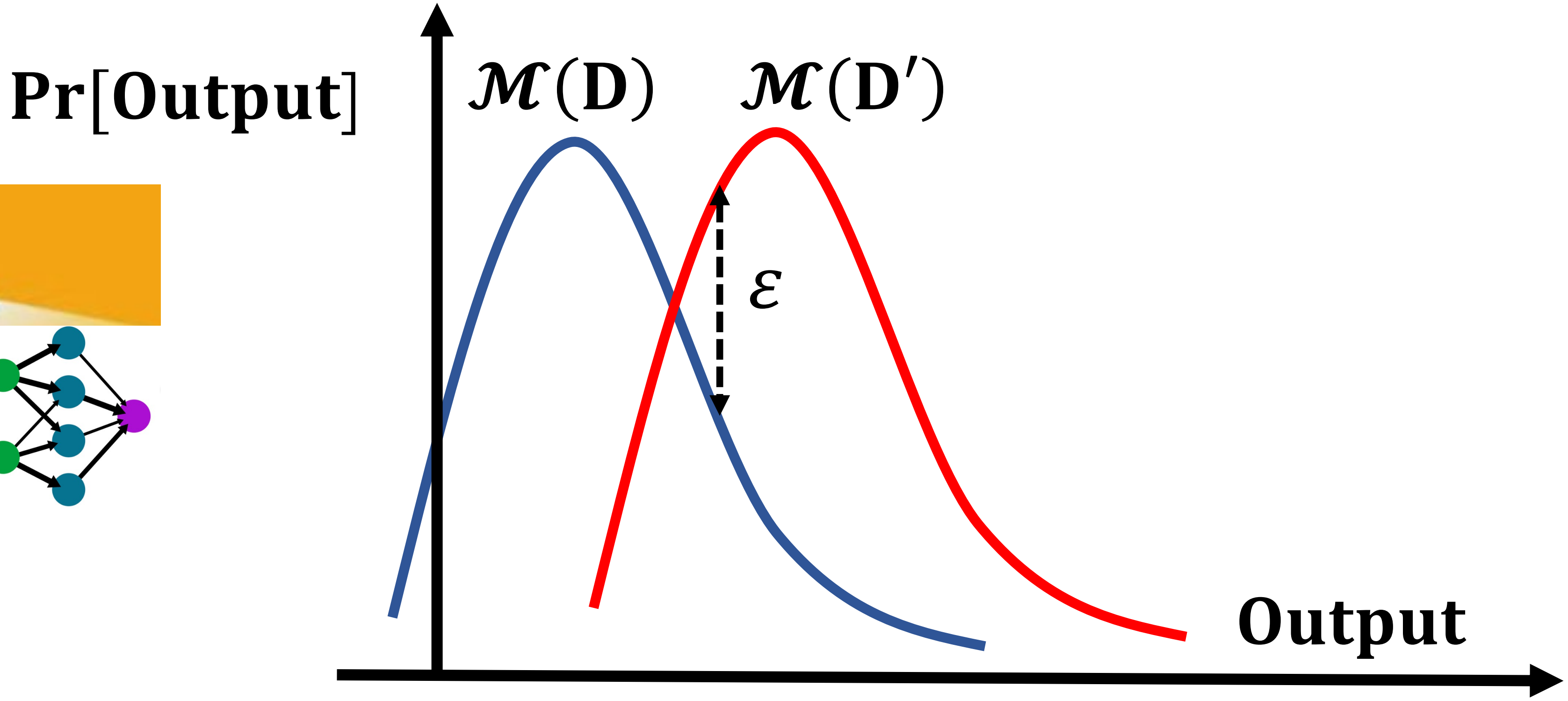


- Observe: gradient updates (stronger threat model than white-box attack)
- Goal: infer the private data

Differential Privacy [DMNS. TCC’ 06]

- For *any* datasets D and D' that differ by *one record*,
  - The output distributions of their transcripts of updates are similar.
  - Similarity is quantified by ε and δ.



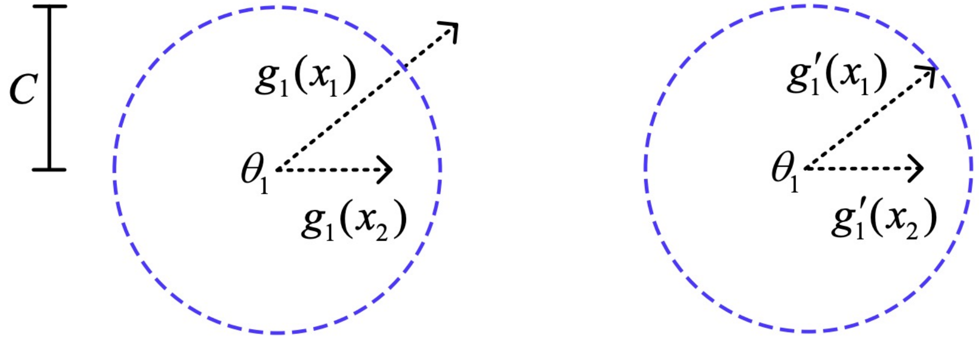


DPSGD [ACGMMTZ. CCS'16]

Step 1. Batch generation: Sample each with probability p.

Step 2. Noise injection:

1) Compute gradient for x_i in the batch B: $g(x_i) = \nabla_{\theta_t} L(\theta_t, x_i)$.
2) Clip gradient



(a) Original gradients. (b) Clipped gradients.

3) Compute gradient sum and add noise $\sum(g(x_i)) + N(0, \sigma^2 C^2 \boldsymbol{I})$.

Step 3. Update parameters with noisy gradient average.

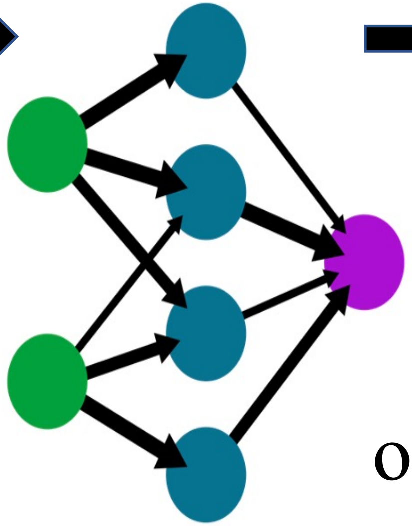
DPIS: Intuition

- Not all records are equally important to training the model [Katharopoulus and Fleuret, ICML' 18].

- Training records with larger gradient norms should be sampled more frequently.

DPIS: Intuition

input



output

Correctly predicted as a “cat”
(smaller gradient norm)

input

output

Wrongly predicted as a “human”
(larger gradient norm)

Choice of Importance

- DPIS samples record x_i with probability proportional to its **importance, denoted as** p_i.

- Natural idea: For record x_i, its importance p_i is **positively-correlated** to its gradient norm $||g(x_i)||$.

- Remember to **scale back** the gradient $g(x_i)$ by factor $1/p_i$, otherwise the gradient sum is a biased estimate.

DPIS: Trade-off

- DPIS differs with DPSGD only in the batch sub-sampling process.

- In DPIS, we let

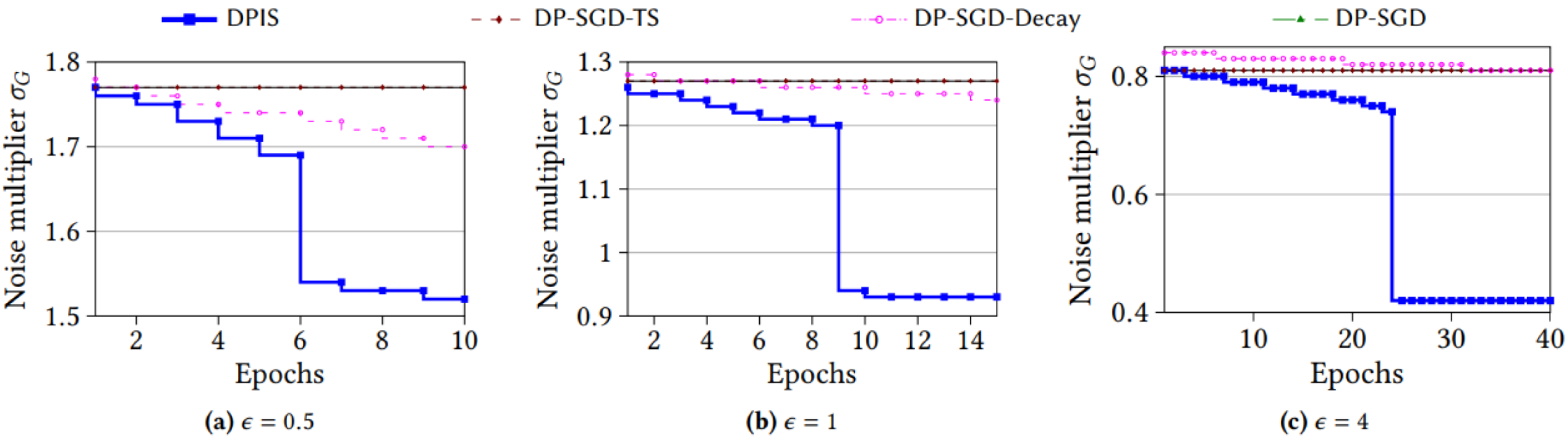
$$p_i = \frac{||g(x_i)||}{K}, \text{ with } \mathrm{K} = \sum_i^N ||g(x_i)||.$$

Importance **Normalizing Factor**

- We show that DPIS improves the accuracy-privacy trade-off at little computation overhead (for the computation of p_i and K).

DPIS: Privacy Analysis

- In MNIST task, fix privacy parameter $\delta = 10^{-5}$, and vary $\varepsilon = 0.5, 1, 4$.
- Compare the noise multipliers for different approaches
  (smaller is better, i.e., larger signal-noise ratio)



Experiments

- Methods:
  - DPIS (ours)
  - DP-SGD
  - DP-SGD with tempered sigmoid functions (DP-SGD-TS)
  - DP-SGD with exponential decay noise (DP-SGD-Decay)

- Results on IMDb (more datasets in our paper):

Method	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$	$\epsilon = 4$
DPIS	**59.1%**	**62.1%**	**65.9%**	**68.3%**	**70.2%**
DP-SGD-TS	57.5%	60.3%	62.8%	65.2%	66.5%
DP-SGD-Decay	56.8%	60.4%	63.6%	66.3%	67.6%
DP-SGD	56.4%	60.3%	63.5%	66.4%	67.4%

(non-DP 79.5%)

Conclusion

- DPIS uses a different sub-sampling approach than DPSGD.

- DPIS samples a record with **importance** proportional to its gradient norm.

- DPIS improves the accuracy while maintaining the same privacy.