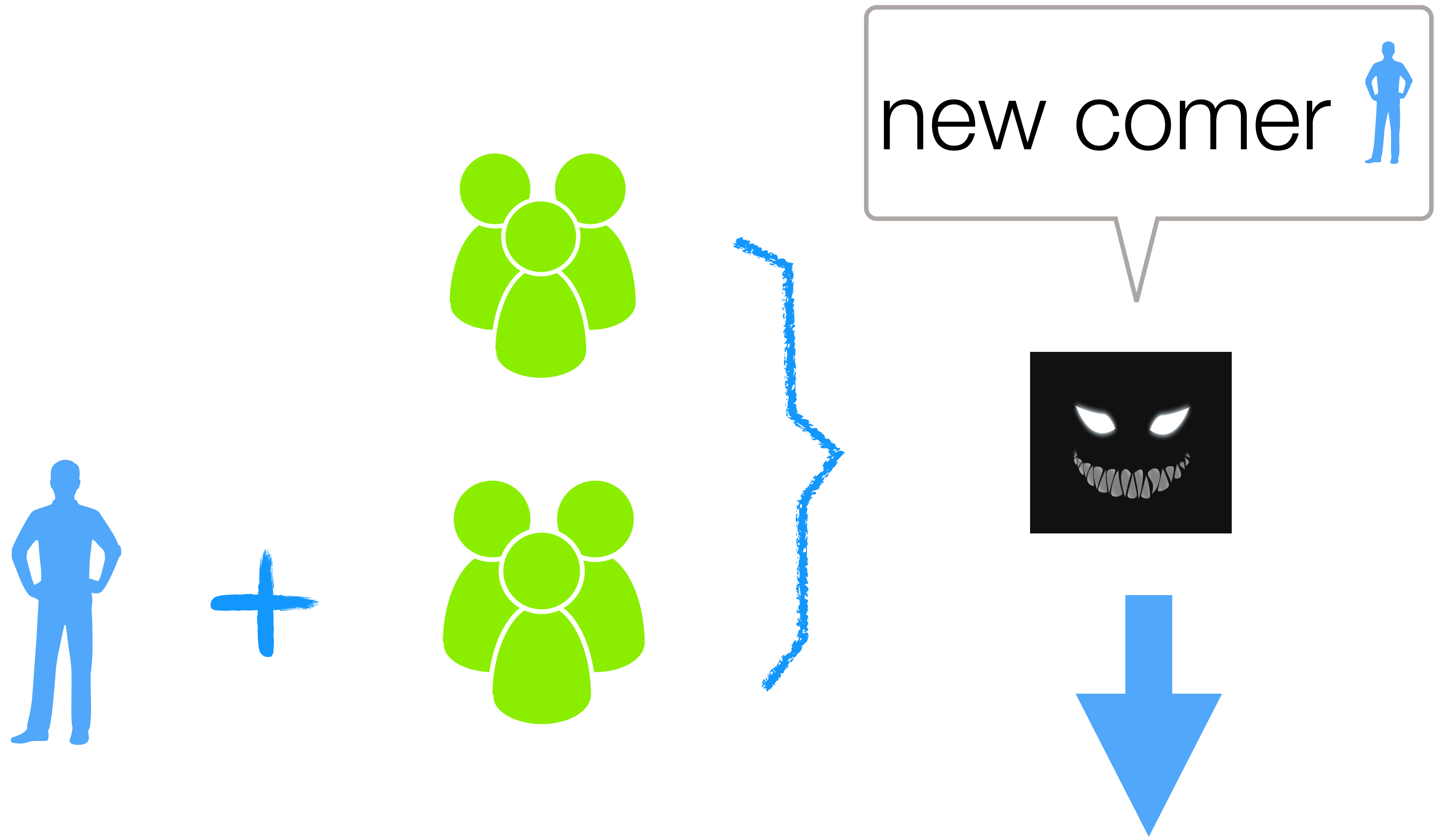
Differentially-Private Deep Learning from an Optimization Perspective

Presenter:
Liyao Xiang
Shanghai Jiao Tong University
4/30/2019

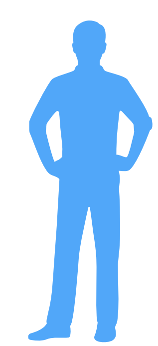
Privacy Threat

- Personal information in big data era
- Is anonymization sufficient to protect user privacy?
- Netflix recommendation challenge: remove personal identity information, replace names with random numbers
- De-anonymize the Netflix database with the public information on IMDb
- De-anonymization even works on partial, distorted, wrong data!

Side Information



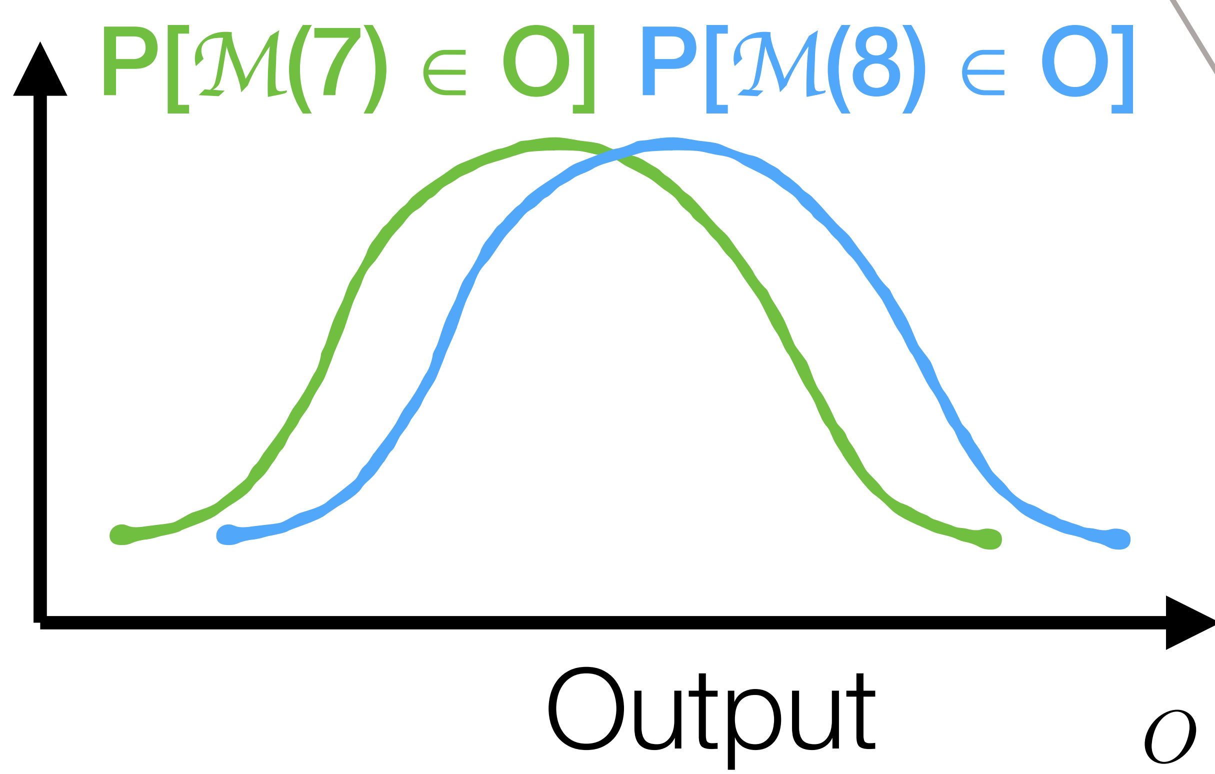
Side information
hurts privacy!



Differential Privacy

adjacent inputs

Constraint: $P[\mathcal{M}(I) \in O] \leq e^{\epsilon} P[\mathcal{M}(I') \in O]$



smaller ϵ indicates higher privacy

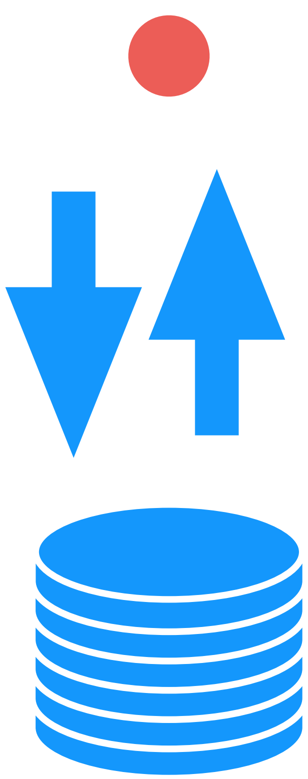
Deep Learning with Differential Privacy

$\theta = (\theta_1, \ldots, \theta_n)$

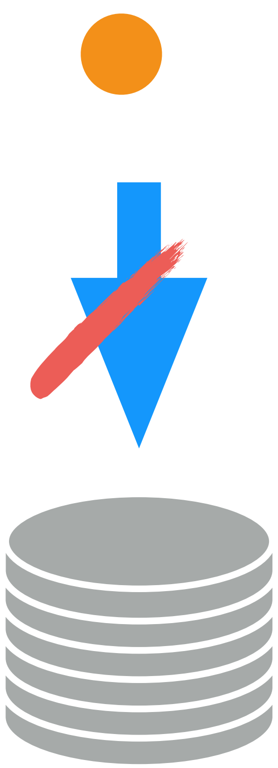
$\vartheta = (\vartheta_1, \ldots, \vartheta_n)$

Perturbation

model tells!

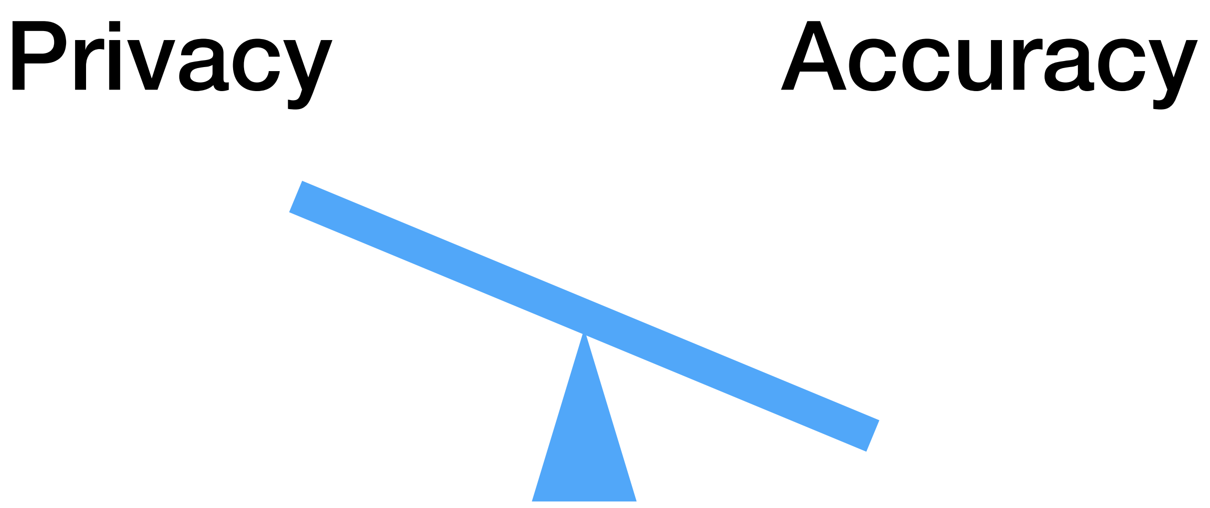


Differentially-private stochastic gradient (DPSGD): add noise to gradient g^t in each iteration of update

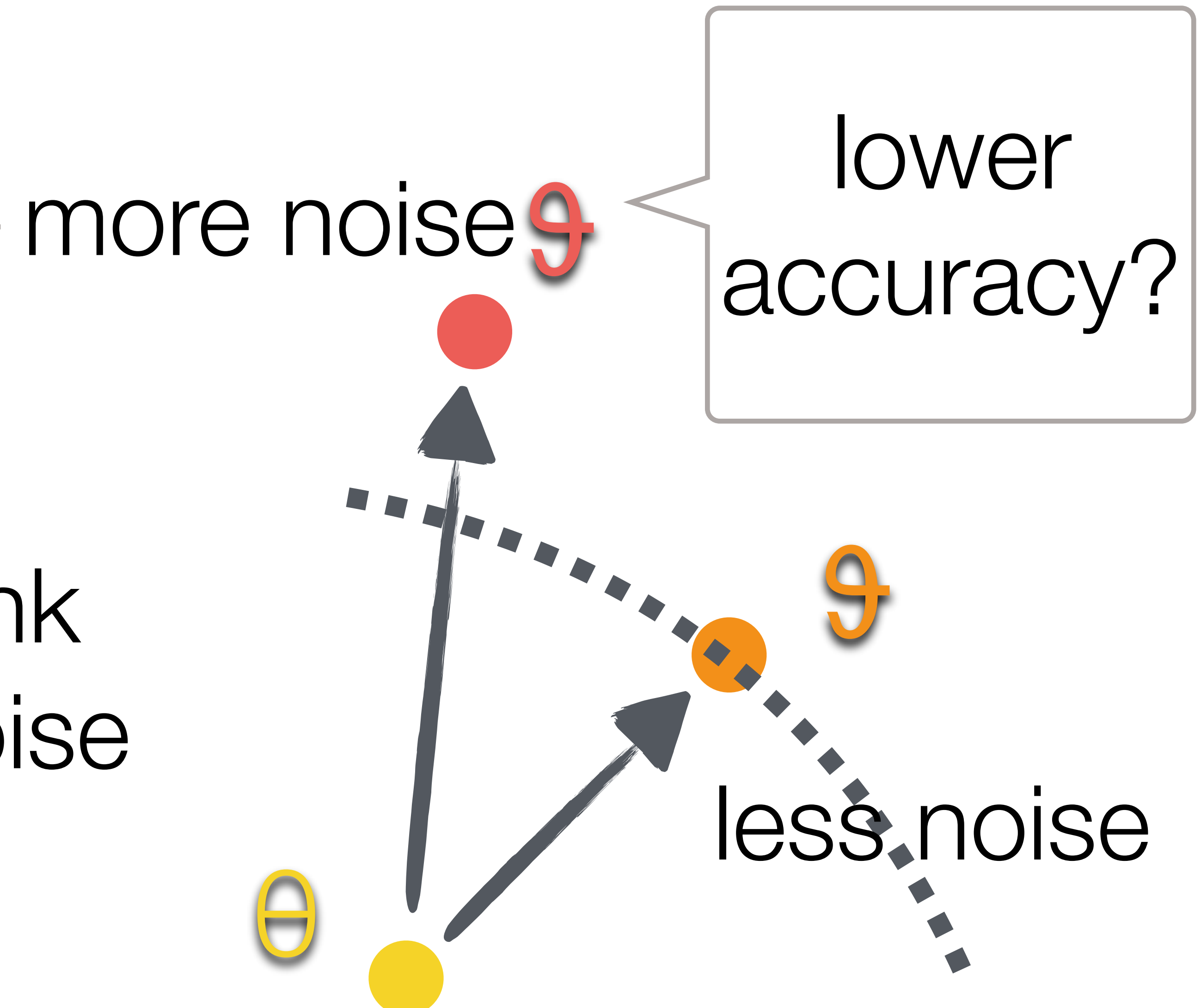


Deep Learning with Differential Privacy

The recent work [Abadi et. al., CCS’ 16] only achieves **~90%** accuracy whereas training w/o privacy reaches over **99%** on MNIST. The result of [Shokri et. al., CCS’ 15] is even worse.

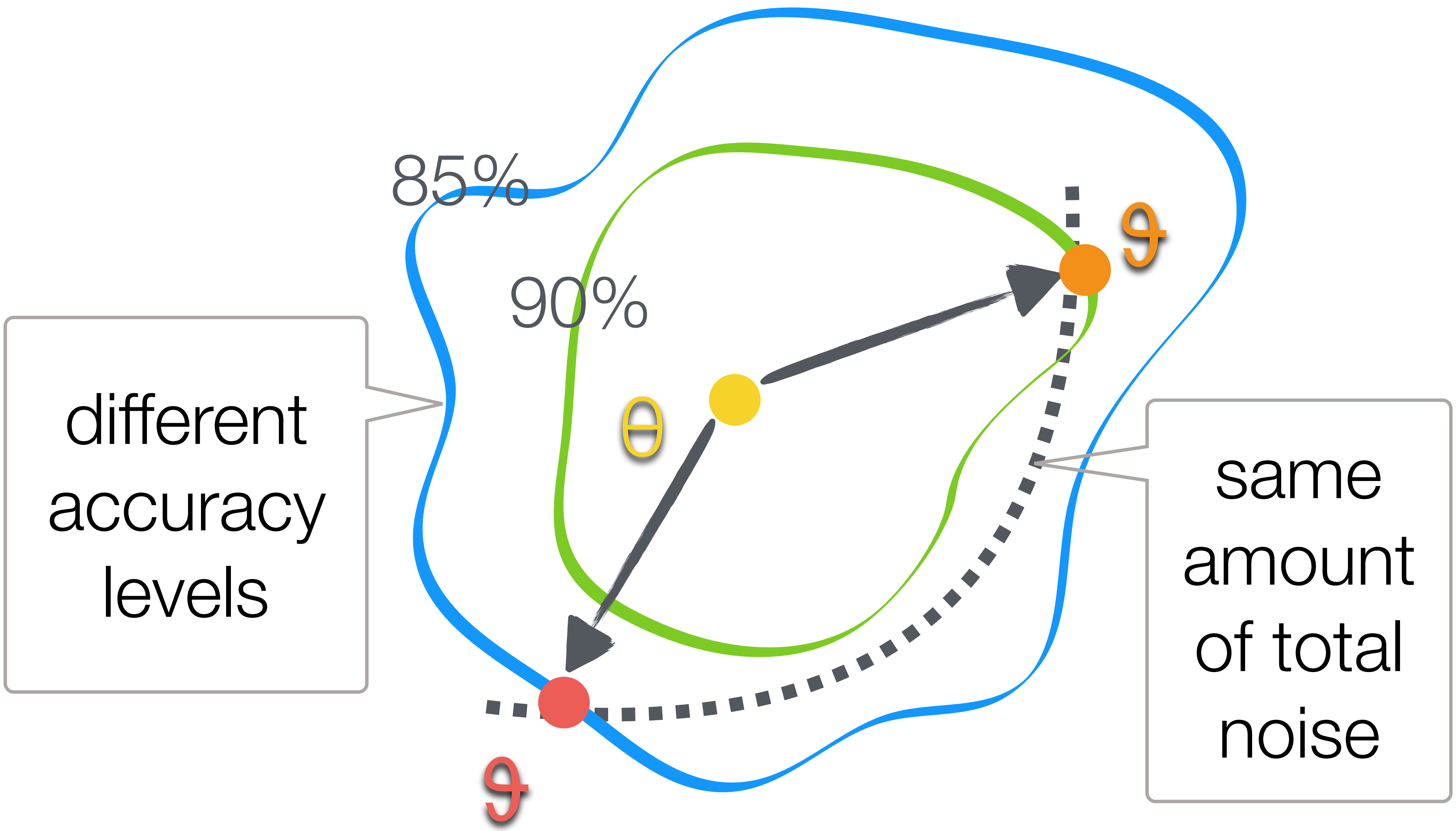


higher privacy level

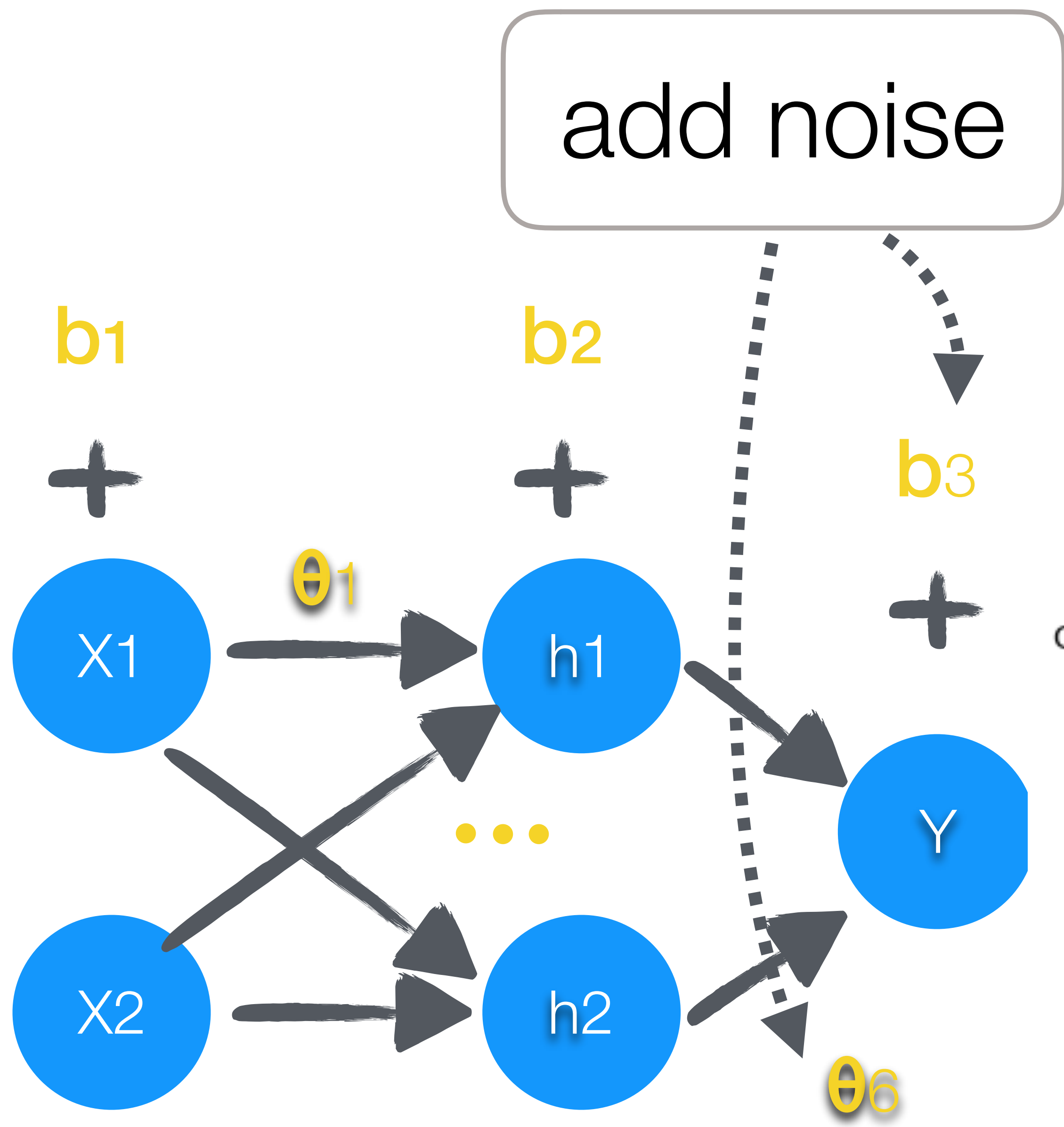


In previous works: link between inserted noise and accuracy is broken

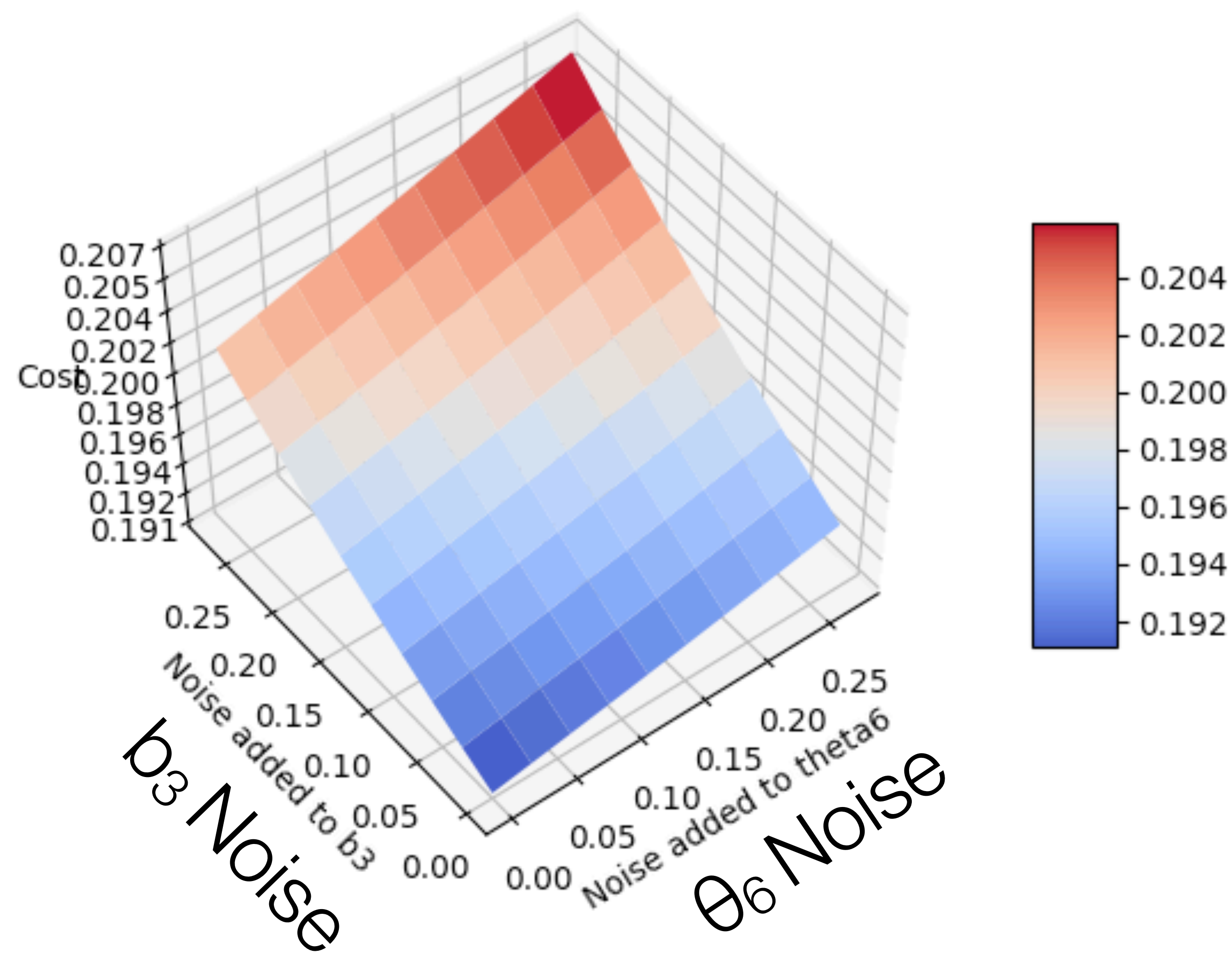
Model Sensitivity



Example



different cost!



Optimized Additive Noise Scheme

- Model sensitivity $\mathbf{w} = (w_1, w_2, \ldots, w_d) \in D^d$: derivative vector of the cost on all training examples w.r.t. all parameters

- To keep the cost minimal, noise should be added to the least sensitive direction of the cost function

- Seek a probability distribution of the noise to minimize the cost as well as to meet differential privacy constraint!

Optimized Additive Noise Scheme

Objective

model sensitivity

distribution of noise

$$\underset{\mathcal{P}}{\text{minimize}} \int_{z_d} \dots \int_{z_1} \langle \mathbf{w}, \mathbf{z} \rangle \mathcal{P}(\mathrm{d}z_1 \dots \mathrm{d}z_d)$$

additive noise

$$w_i = \frac{\partial C}{\partial \theta_i} > 0$$

cost increases as θ_i increases ⇒ pushes z_i to a direction where $z_i < 0$

$$\frac{\partial C}{\partial \theta_i} > \frac{\partial C}{\partial \theta_j} > 0$$

cost is more sensitive to changes of θ_i than θ_j ⇒ less noise should be added to θi

Optimized Additive Noise Scheme

Constraint

global sensitivity on adjacent inputs: $\alpha = \sup_{\forall \mathbf{X}, \mathbf{X}' \text{ s.t. } d(\mathbf{X}, \mathbf{X}')=1} \|\mathbf{g}^t - \mathbf{g}'^t\|$

training datasets differ by a single instance

L2-norm between the gradients

$$\Pr[\mathcal{M}(\mathbf{g}^t) \in \mathcal{O}] \leq e^{\epsilon} \Pr[\mathcal{M}(\mathbf{g}'^t) \in \mathcal{O}]$$

$$\Rightarrow \Pr[\mathbf{g}^t + \mathbf{z} \in \mathcal{O}] \leq e^{\epsilon} \Pr[\mathbf{g}'^t + \mathbf{z} \in \mathcal{O}]$$

$$\Rightarrow \Pr[\mathbf{z} \in \mathcal{O} - \mathbf{g}^t] \leq e^{\epsilon} \Pr[\mathbf{z} \in \mathcal{O} - \mathbf{g}'^t]$$

$$\Rightarrow \Pr[\mathbf{z} \in \mathcal{O}'] \leq e^{\epsilon} \Pr[\mathbf{z} \in \mathcal{O}' + \mathbf{g}^t - \mathbf{g}'^t]$$

Optimized Additive Noise Scheme

$$\underset{\mathcal{P}}{\text{minimize}} \int_{z_d} \ldots \int_{z_1} \langle \mathbf{w}, \mathbf{z} \rangle \mathcal{P}(\mathrm{d}z_1 \ldots \mathrm{d}z_d)$$

$$\text{s.t. } \Pr[\mathbf{z} \in O'] \leq e^{\epsilon} \Pr[\mathbf{z} \in O' + \Delta]$$

$$\forall O' \subseteq \mathbb{R}^d, ||\Delta|| \leq \alpha$$

$$\underset{p}{\text{minimize}} \int_{\mathbf{z} \in \mathbb{R}^d} \|\mathbf{w} \circ \mathbf{z}\|_1 p(\mathbf{z}) \mathrm{d}\mathbf{z}$$

$$\text{s.t. } \ln \frac{p(\mathbf{z})}{p(\mathbf{z} + \Delta)} \leq \epsilon, \; \forall ||\Delta|| \leq \alpha, \Delta \in \mathbb{R}^d.$$

Composition

- So far, we only show how to provide privacy guarantee in a single iteration of update
- In practice, SGD takes many iterations until convergence
- Iterative computation exposes the training data multiple times, degrading privacy level!
- Our solution: Advanced composition theorem for differential privacy + privacy amplification by sampling

Optimized Additive Noise Mechanism

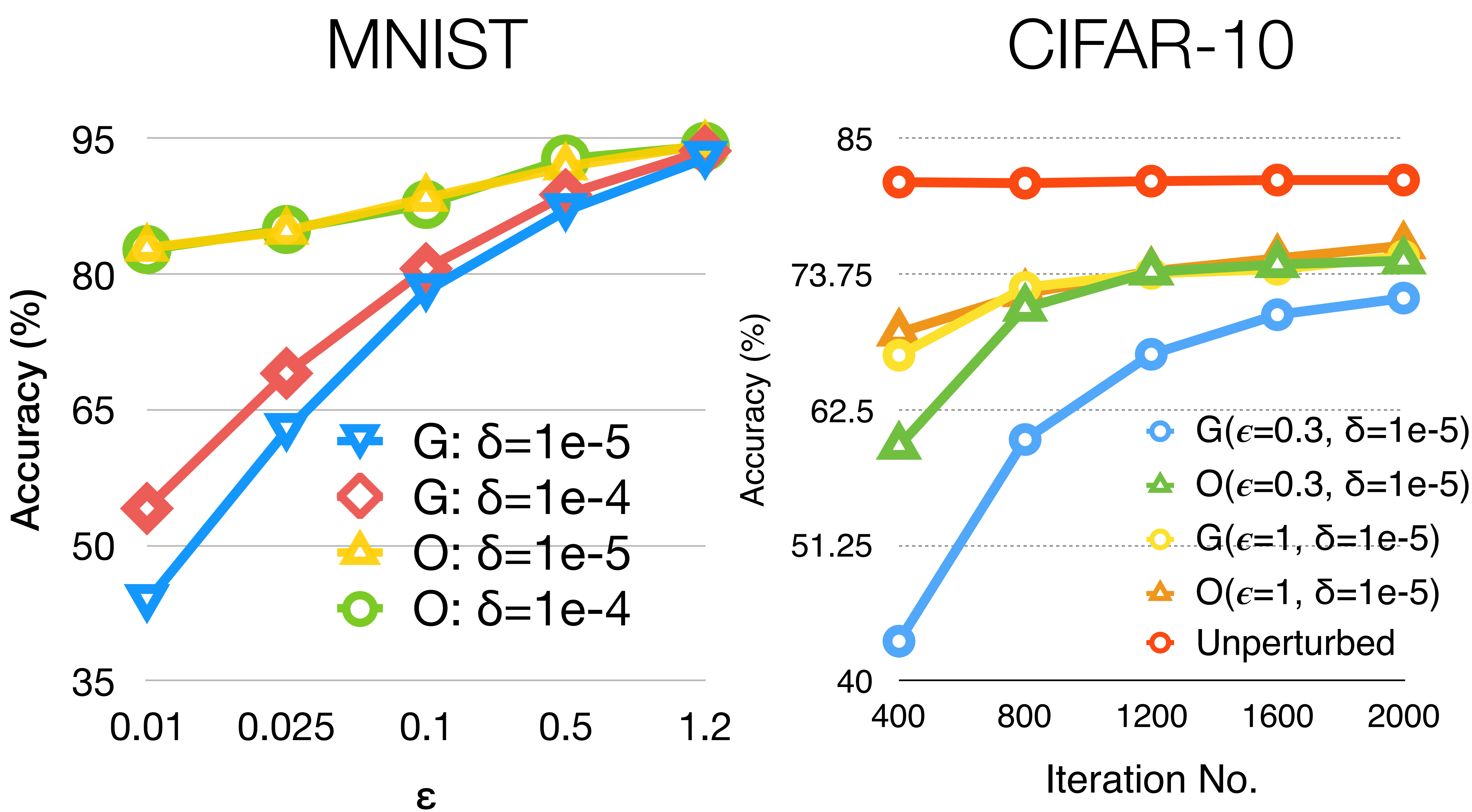
1. Compute per-iteration privacy parameters according to composition theorem
2. For each iteration
    1. Compute model sensitivity **w**
    2. Solve the optimization problem to find noise distribution
    3. Sample a noise
    4. For each batch of training data: Compute and clip the gradient by global sensitivity
    5. Compute the average gradient for the batch
    6. Add noise to the average gradient
    7. Update model parameters

Implementation

Implement optimized noise generator (ours) and Gaussian noise generator (the state-of-the-art, Abadi et. al.) on Keras and Tensorflow

Problem: computational challenges due to high dimensionality

- ☑ Solving the optimization problem using GPU operations
- ☑ Numpy noise generator



Our scheme achieves higher accuracy over [Abadi CCS' 16] under the same privacy guarantee

Thank you!