Private Adaptive Gradient Methods for Convex Optimization

Alireza Fallah*
MIT

Joint work with:

Hilal Asi*
Stanford

John Duchi
Stanford

Omid Javidbakht
Apple

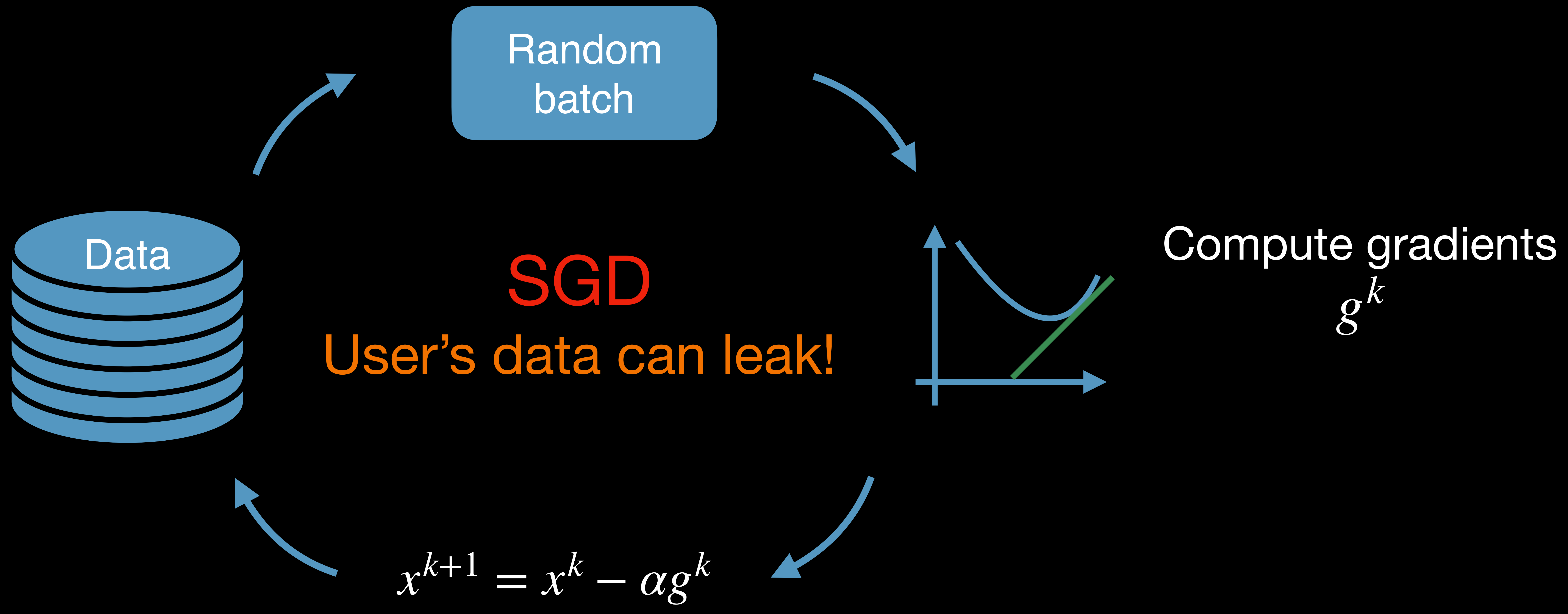
Kunal Talwar
Apple

*** Work done while interning at Apple.**

ICML 2021

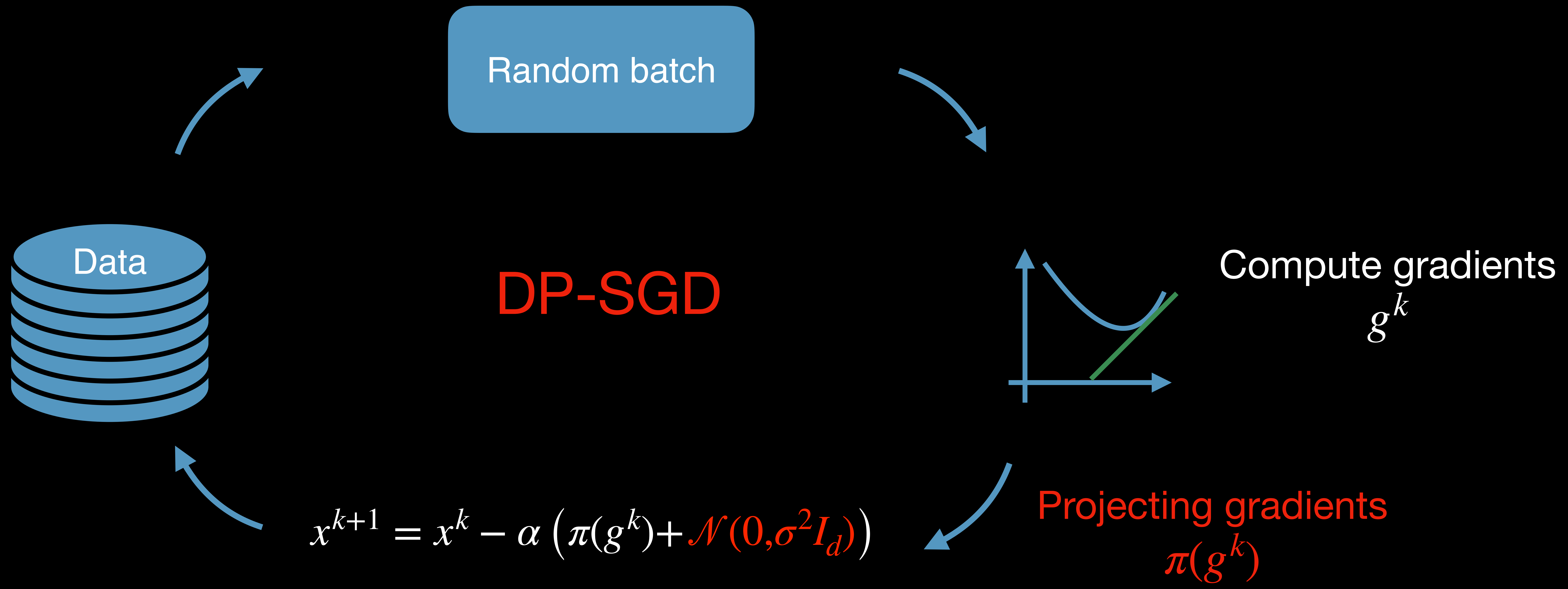
Stochastic Gradient Methods

- Goal: solving $\min_{x \in \mathscr{X}} f(x) := \frac{1}{n} \sum_{i=1}^{n} F(x, z_i)$ privately ($\mathscr{X} \subset \mathbb{R}^d$ is a convex set).
- F is a *convex (possibly non-smooth)* function & $\{z_1, \cdots, z_n\}$ are n datapoint.



Private Stochastic Gradient Methods

- A randomized algorithm $\mathscr{M}$ is (ε, δ) differentially private (DP) if for all neighboring datasets $\mathscr{S}, \mathscr{S}'$ we have: $\mathbb{P}(\mathscr{M}(\mathscr{S}) \in O) \leq e^{\varepsilon}\, \mathbb{P}(\mathscr{M}(\mathscr{S}') \in O) + \delta$ (for any open interval O).

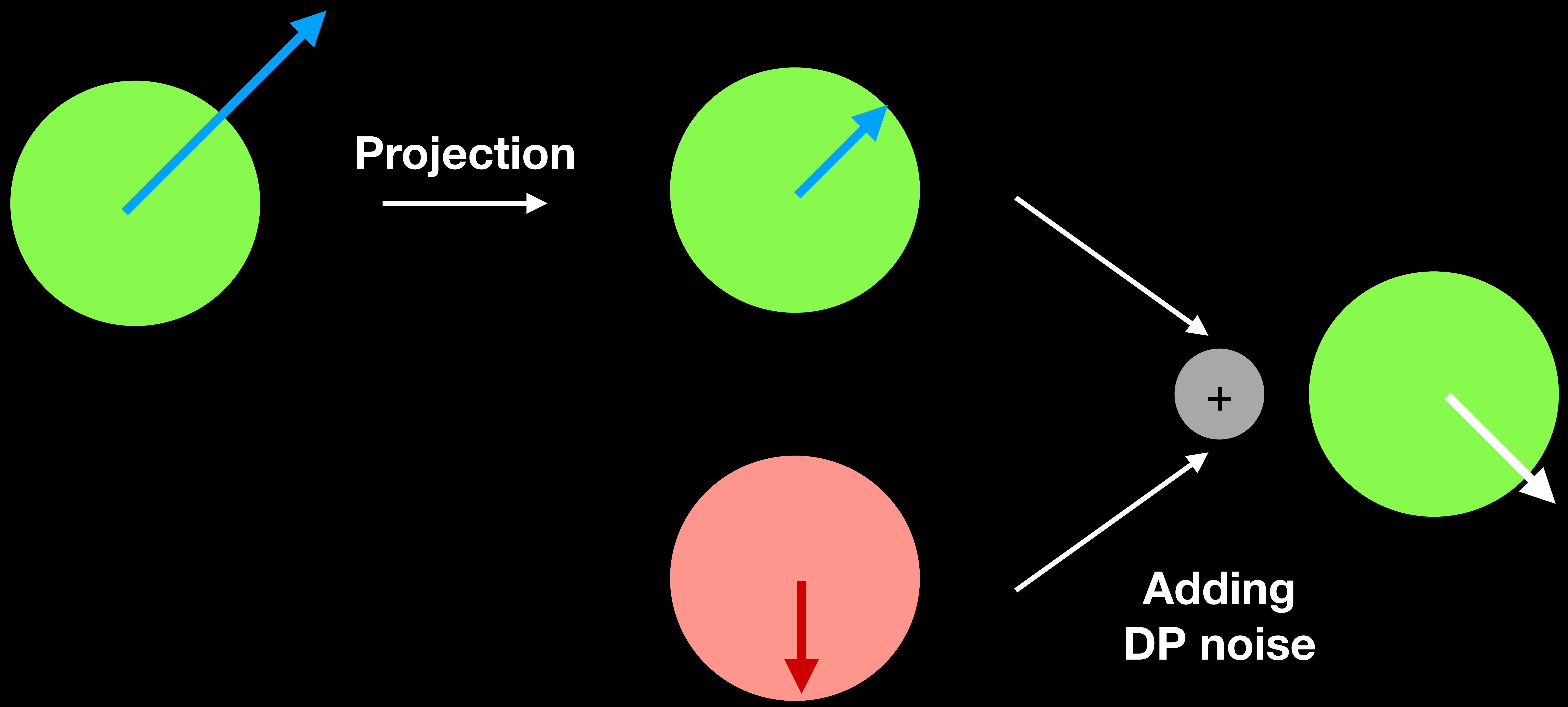


Private Adaptive Gradient Methods

- In this work, we propose two novel **private** **adaptive** gradient methods.
- Our algorithms are ***adaptive*** in two aspects:
  1. The projection (and hence the added noise) adapts to the underlying geometry of gradients.
  2. We use adaptive optimization methods to further exploit the underlying geometry of the problem.
- We explain these two aspects separately.

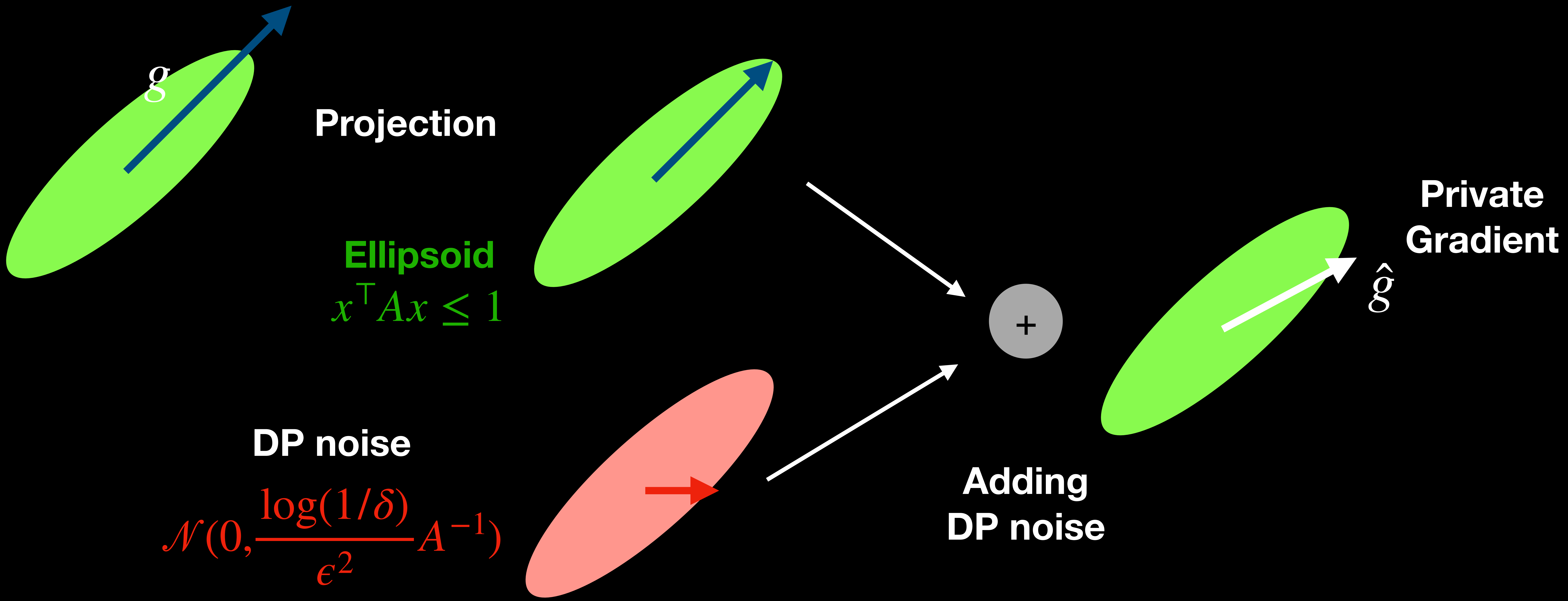
Projection in Original DP-SGD

- Original DP-SGD projects gradients to a d dimensional ball.



Projection in Our DP Algorithms

- In many applications gradients lie in certain geometries, e.g., they are sparse.
- We could take advantage of this geometry in both projection and adding DP noise.



PASAN & PAGAN

Private Adaptive SGD/AdaGrad with Adaptive Noise

- We further leverage gradient geometry by choosing adaptive learning rates.
- Our proposed methods (PASAN & PAGAN) use the aforementioned projection.
- **PASAN**: Private SGD with **adaptive scalar** learning rates: $\alpha_k = \alpha(\sum_{i=1}^{k} ||\hat{g}^i||^2)^{-1/2}$
- **PAGAN**: Private AdaGrad, i.e., using **adaptive diagonal matrices** as learning rates:

$$\alpha_k = \alpha \left(\text{diag}(\sum_{i=0}^{k} \hat{g}^i \hat{g}^{i\top}) \right)^{-1/2}$$

Convergence of PASAN and PAGAN

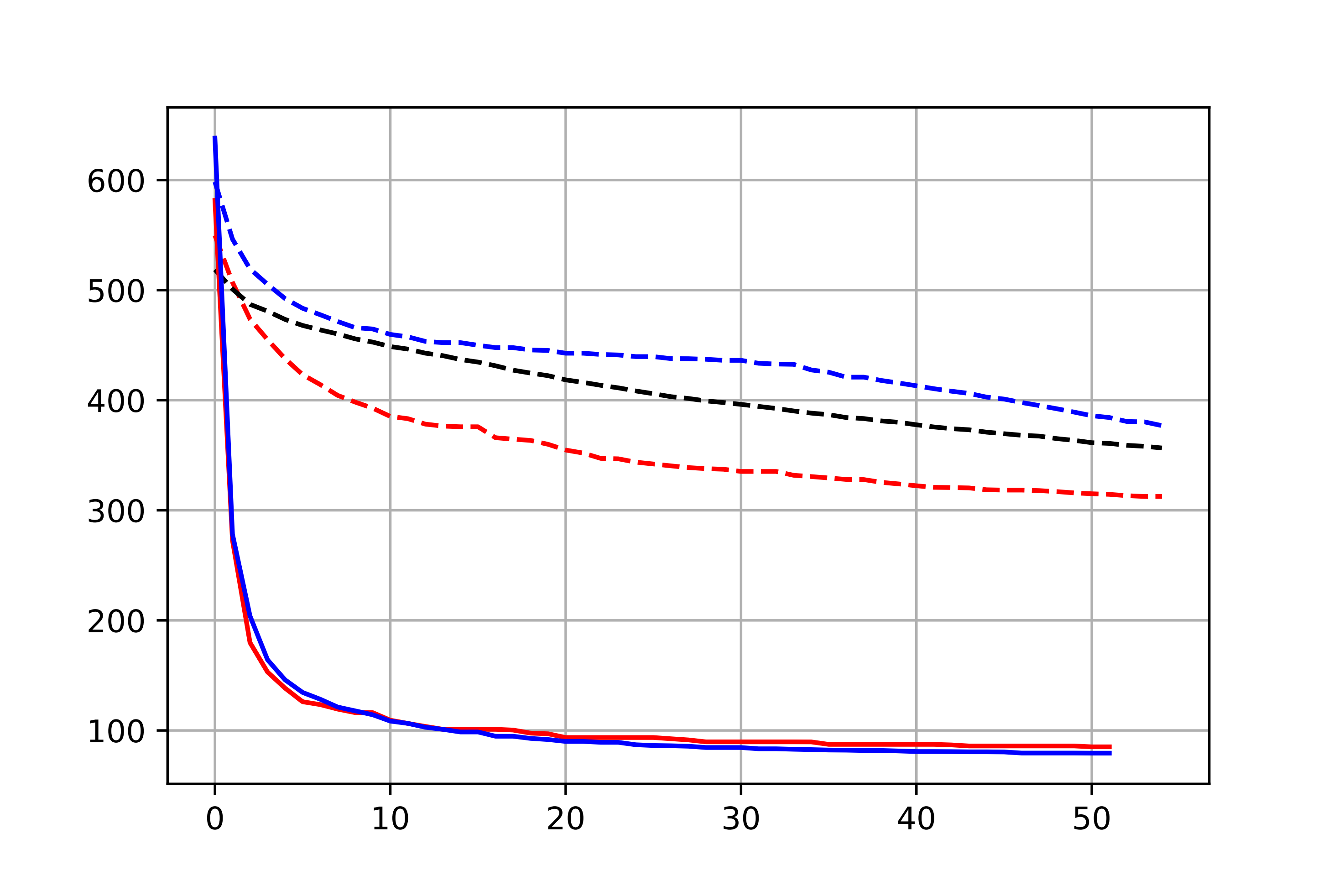
- We provide upper and lower bounds for the convergence of PASAN and PAGAN.
- Main assumption to capture the underlying geometry of gradients:
  - The l_p norm of Lipschitz constant with respect to a matrix norm is bounded, i.e.:

$$\mathbb{E}_z[||\nabla F(x,z)||_C^p] \leq G^p$$

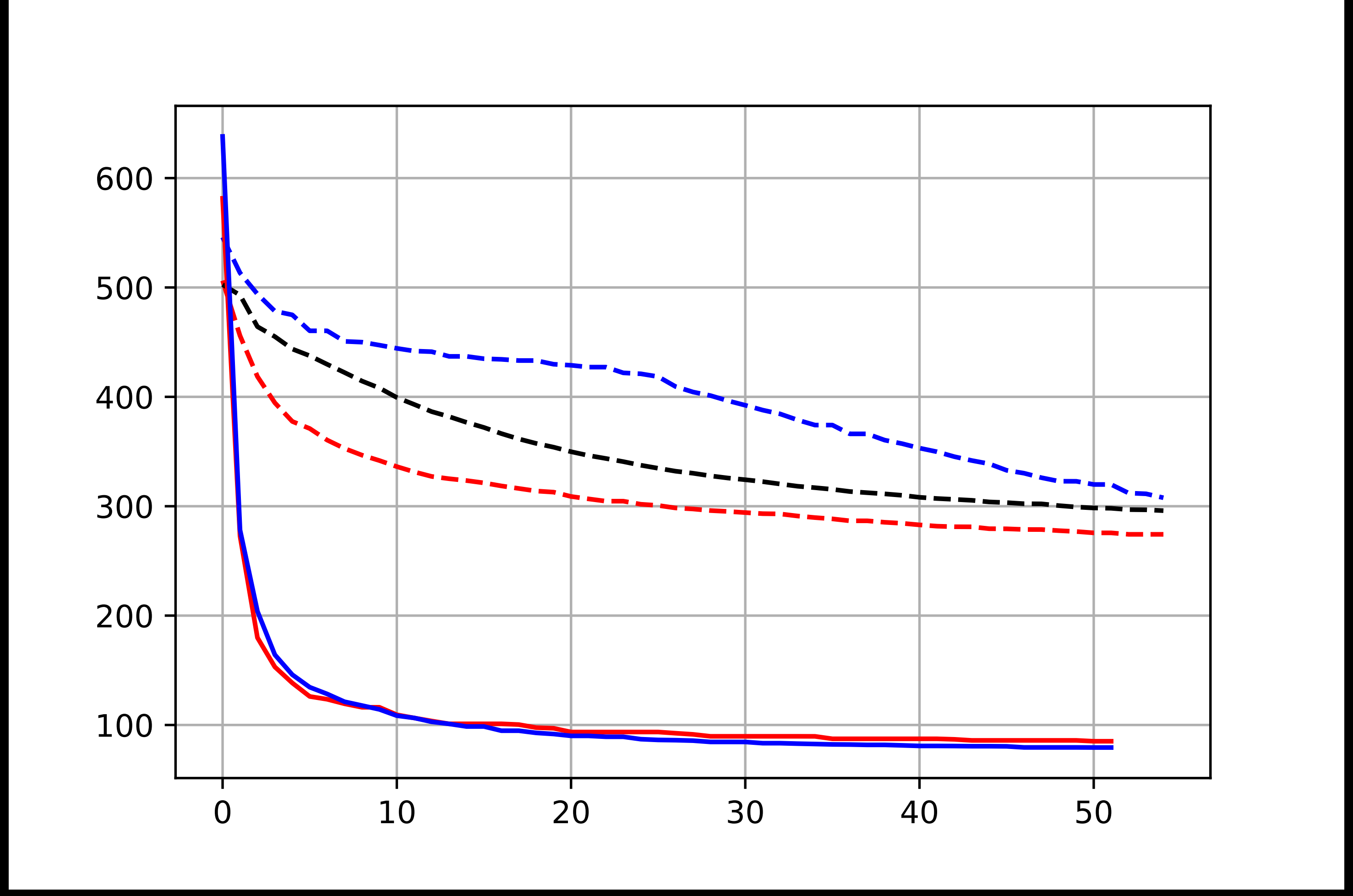
- Our results show that, in certain cases where gradients are sparse across coordinates, PAGAN improves dimension dependence up to a factor of $\sqrt{d}/\log(d)$!

Experiments

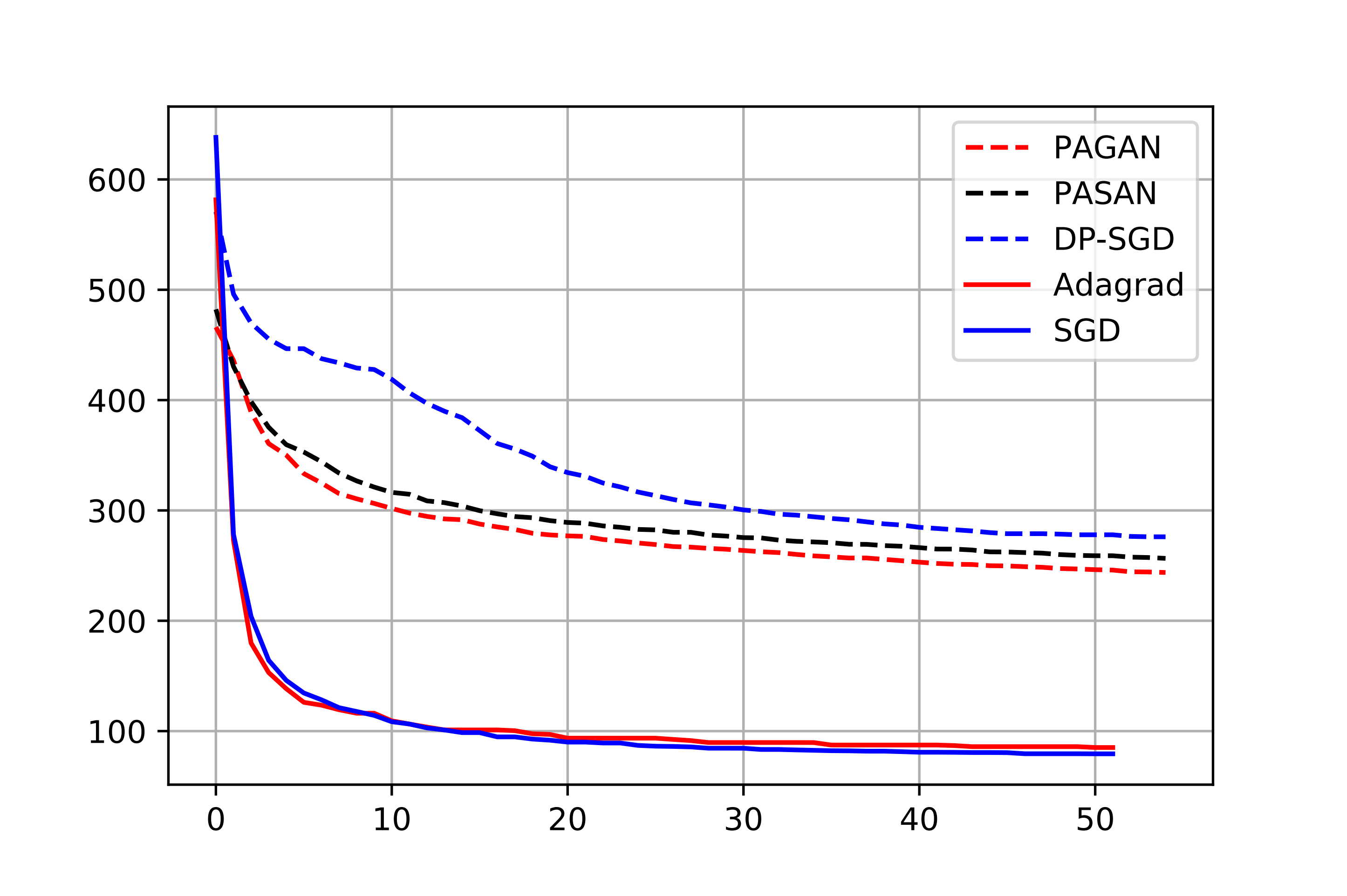
- We train an LSTM model over WikiText-2 dataset (details in paper.)
- We report minimum validation perplexity vs. training rounds (7 epochs).



$\varepsilon = 0.5$



$\varepsilon = 1$



$\varepsilon = 3$

Minimum validation perplexity

Experiments (Cont.)

Test Perpelexity	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 3$
DP-SGD	349.1	287.81	240.32
PASAN	332.52	274.63	238.87
PAGAN	291.41	253.41	224.82

- Additional experiments in convex setting in paper.

Stop by and check our poster for further and more detailed results!

Thanks!