Stochastic adversarial noise in non-smooth convex federated optimization

Mashalov Nikita
Zernova Anna
Mosoyan Kamo

Email: mashalov.ne@phystech.edu

Presentation plan

- task introduction
- proof results
- numerical experiment
- your questions :)

Problem introduction

Optimization

General problem of optimization is formulated as

$$\min_{x \in D} f(x)$$

Deep learning interpretation

$$\min_{x \in D} f(x) = \min_{x \in D} \mathbf{E}_{\xi \sim \pi} f(x, \xi)$$

ξ - samples from training set defined by probability distribution π

x - weights of model, which are restricted to set D

$f(x, \xi)$ - defines model inference

Oracle concept

Optimized goal function ***f(x)*** is known only by oracle

Oracle in deep learning

- Oracle calls = model evaluation
- Neural nets requires lot’s of computation
- Method with less calls of oracle can learn model much faster :)

Adversarial noise types

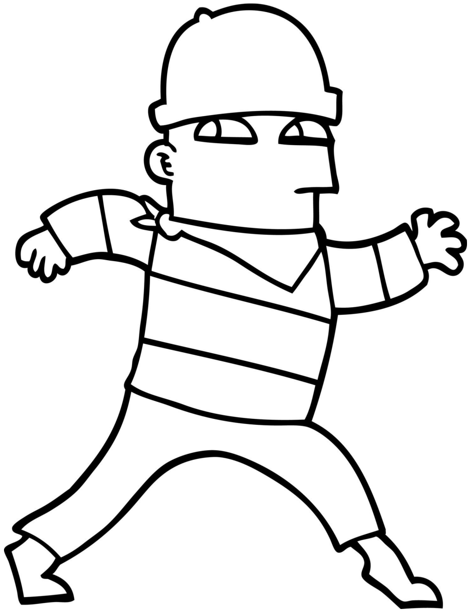
Following types of constraints on noise were introduced:

- Deterministic, coordinate (weight) dependant and restricted value

$$\delta(x), |\delta(x)| < \Delta$$

- Stochastic noise from unknown distribution with limited second moment

$$\delta \sim p(\delta), \mathbf{E}_\delta^2 < \Delta^2$$



https://arxiv.org/pdf/2304.07861.pdf

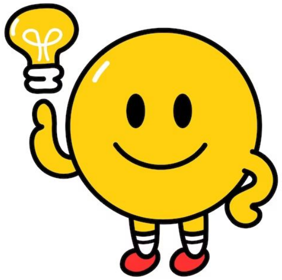
Zero gradients methods

Curse of dimensionality

Non-gradients methods requires $\approx \sqrt{d}$ times more Iterations than gradients ones, where d is number of model parameters

Non gradient methods are effective only in low-dimensionality methods

Is that really bad?



Not always :)

Recent advances in deep learning shows that even low-dimension model augmentations can be effective

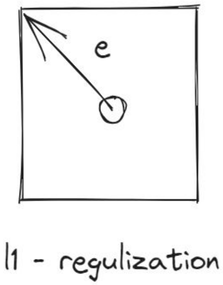
- LoRA
- Compression operators

Lora: https://arxiv.org/abs/2106.09685
Compression operators: https://arxiv.org/pdf/2206.09446

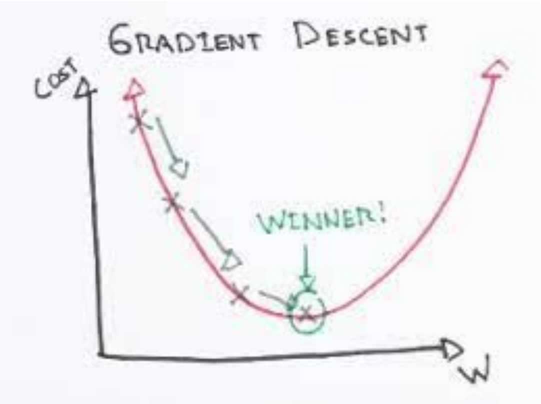
Zero gradients roadmap

Optimization cycle

Oracle



Gradient Approximation



Gradient algorithm

Next step

Assumptions

Following assumptions were introduced:

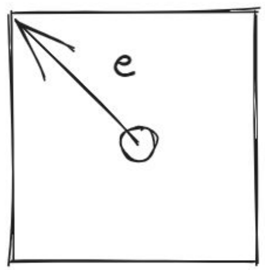
- Lipschitz continuity

$$|f(y,\xi) - f(x,\xi)| \le M(\xi)\|x - y\|_p$$

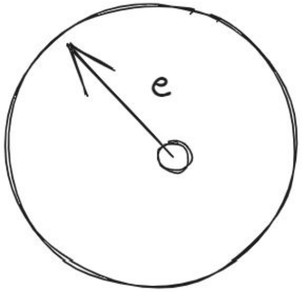
- Convexity on optimization set

$$Q_\gamma = Q + B_p^d(\gamma)$$

l1 / l2 regulization: stochastic approximation of gradient



l1 - regulization



l2 - regulization

$$\nabla f_\gamma(x,\xi,e) = \frac{d}{2\gamma}(f_\delta(x+\gamma e,\xi) - f_\delta(x-\gamma e,\xi))\mathrm{sign}(e).$$

$$\nabla f_\gamma(x,\xi,e) = \frac{d}{2\gamma}(f_\delta(x+\gamma e,\xi) - f_\delta(x-\gamma e,\xi))e.$$

More general overview can be obtained in:
https://arxiv.org/pdf/2211.10783.pdf

These regularization are good :)

Optimization with following smoothing parameters allows to find solution with required error tolerance

l1- regularization

$$\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$$

l2 - regularization

$$\gamma = \frac{\varepsilon}{2M_2}$$

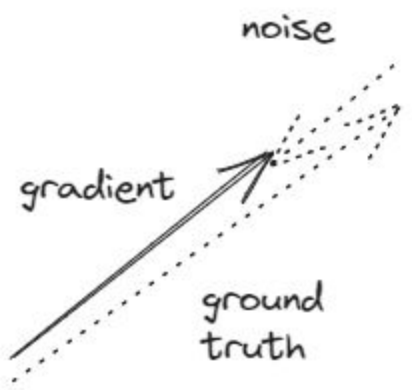
More general overview can be obtained in:
https://arxiv.org/pdf/2211.10783.pdf

Noise affection

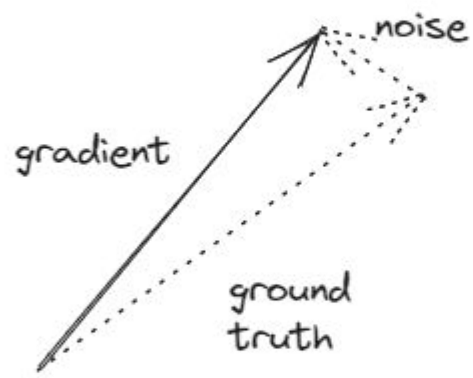
How noise affects gradient optimization?

Noise affection can be decomposed to two instances

Gradient scale = variance



Gradient orientation = bias



Stochastic noise regularized gradient estimation

Gradient estimation is unbiased

$$\mathbf{E}_{e,\xi}[\nabla f_\gamma(x,e,\xi)] = \nabla f_\gamma$$

Gradient variance

l1- regularization

$$\kappa(p,d)\left(M_2^2 + \frac{d^2\Delta^2}{12(1+\sqrt{2})^2\gamma^2}\right),$$

l2 - regularization

$$\kappa(p,d)\left(dM_2^2 + \frac{d^2\Delta^2}{\sqrt{2}\gamma^2}\right),$$

Following results are compilation of awesome paper:
https://arxiv.org/pdf/2304.07861.pdf

Majorization (l1-example)

Recall of optimal smoothing

$$\gamma = \frac{\sqrt{d}\varepsilon}{4M_2}$$

$$\kappa(p,d)\left(M_2^2 + \frac{d^2\Delta^2}{12(1+\sqrt{2})^2\gamma^2}\right),$$

Approximation noise $\gg$ Adversarial noise

$$\boxed{\Delta \lesssim \frac{\varepsilon}{\sqrt{d}}}$$

Following results are compilation of awesome paper:
https://arxiv.org/pdf/2304.07861.pdf

Results

Adversarial noise level can be relaxed for stochastic case

<u>Deterministic noise</u>

$$\boxed{\Delta \lesssim \frac{\varepsilon^2}{\sqrt{d}}}$$

<u>Stochastic noise</u>

$$\boxed{\Delta \lesssim \frac{\varepsilon}{\sqrt{d}}}$$

Obtained win is result of absence of bias in gradient estimation

Numerical experiment

Numerical experiment settings

$$f(x,\xi) = \sqrt{\left(< x, \xi > - \sum_{i=1}^{d} \xi_i \right)^2} + \|x\|_\infty + \delta$$

$x \in R^d$

$\xi_i \sim U[0,1]$

$\delta \sim \mathbf{N}(\mathbf{0}, \sigma^\mathbf{2})$

Linear regression

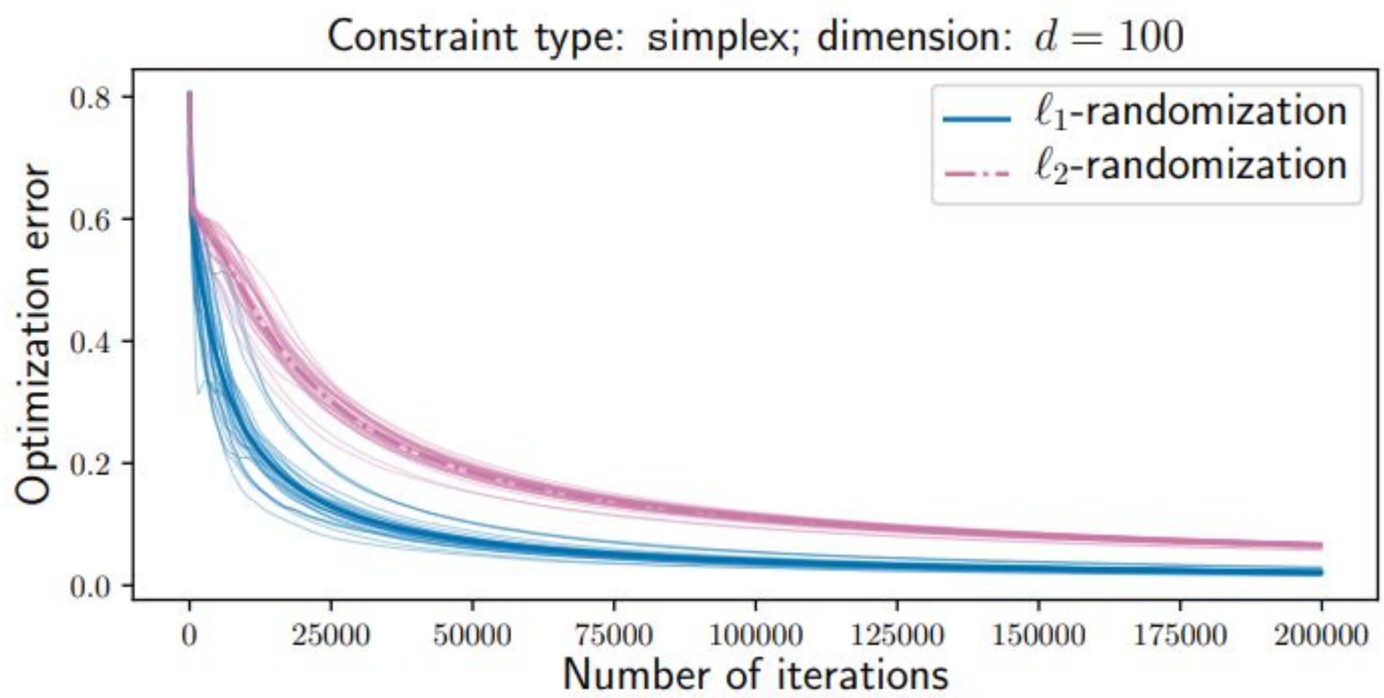
Non flexible regularization

Noise

You can reproduce experiment by visiting github repository

https://github.com/NMashalov/FederationLearning

Numerical experiment settings



Similar results were obtained in: https://arxiv.org/pdf/2205.13910.pdf

Thank you for attention!

Questions section ?