From low probability to high confidence in stochastic convex optimization

Dmitriy Drusvyatskiy
Mathematics, University of Washington

Joint work with D. Davis (Cornell), L. Xiao (MSR), J. Zhang (UMTC)

IFDS Brown-Bag 2020

Problem.

$$\min_x \; F(x) = \mathbb{E}_{z \sim P}[f(x, z)]$$

where $f(\cdot, z)$ are L-smooth and μ-strongly convex. Condition number: $\kappa = \frac{L}{\mu}$.

Problem.

$$\min_x \; F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

where $f(\cdot, z)$ are L-smooth and μ-strongly convex. Condition number: $\kappa = \frac{L}{\mu}$.

Sample efficiency. Typical algorithms find x satisfying

$$\mathbb{E}[F(x) - F^*] \le \epsilon \qquad \text{using} \qquad O\left(\frac{1}{\epsilon}\right) \text{ samples,}$$

(**eg**: Empirical risk minimization, variants of stochastic gradient, ...)

Problem.

$$\min_x \; F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

where $f(\cdot, z)$ are L-smooth and μ-strongly convex. Condition number: $\kappa = \frac{L}{\mu}$.

Sample efficiency. Typical algorithms find x satisfying

$$\mathbb{E}[F(x) - F^*] \leq \epsilon \qquad \text{using} \qquad O\left(\frac{1}{\epsilon}\right) \text{ samples,}$$

(**eg**: Empirical risk minimization, variants of stochastic gradient, ...)

Therefore (Markov)

$$\mathbb{P}\left(F(x) - F^* \leq \epsilon\right) \geq 1 - p \qquad \text{using} \qquad O\left(\frac{1}{\epsilon p}\right) \text{ samples.}$$

Problem.

$$\min_x \; F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

where $f(\cdot, z)$ are L-smooth and μ-strongly convex. Condition number: $\kappa = \frac{L}{\mu}$.

Sample efficiency. Typical algorithms find x satisfying

$$\mathbb{E}[F(x) - F^*] \leq \epsilon \qquad \text{using} \qquad O\left(\frac{1}{\epsilon}\right) \text{ samples,}$$

(**eg**: Empirical risk minimization, variants of stochastic gradient, ...)

Therefore (Markov)

$$\mathbb{P}\left(F(x) - F^* \leq \epsilon\right) \geq 1 - p \qquad \text{using} \qquad O\left(\frac{1}{\epsilon p}\right) \text{ samples.}$$

Question.

What is the true overhead cost of high confidence?

$$\mathbb{P}\left(F(x) - F^* \leq \epsilon\right) \geq 1 - p$$

Problem.

$$\min_x \; F(x) = \mathbb{E}_{z \sim P}[f(x, z)]$$

where $f(\cdot, z)$ are L-smooth and μ-strongly convex. Condition number: $\kappa = \frac{L}{\mu}$.

Sample efficiency. Typical algorithms find x satisfying

$$\mathbb{E}[F(x) - F^*] \leq \epsilon \qquad \text{using} \qquad O\left(\frac{1}{\epsilon}\right) \text{ samples,}$$

(**eg**: Empirical risk minimization, variants of stochastic gradient, ...)

Therefore (Markov)

$$\mathbb{P}\left(F(x) - F^* \leq \epsilon\right) \geq 1 - p \qquad \text{using} \qquad O\left(\frac{1}{\epsilon p}\right) \text{ samples.}$$

Question.

What is the true overhead cost of high confidence?

$$\mathbb{P}\left(F(x) - F^* \leq \epsilon\right) \geq 1 - p$$

Wishful thinking: Might $O\left(\frac{1}{\epsilon}\log(\frac{1}{p})\right)$ samples suffice?

A formal question

$$\min_{x} \; F(x) = \mathbb{E}_{z \sim P}[f(x,z)]$$

Minimization Oracle: Suppose $\mathcal{M}_F(\epsilon)$ returns x_ϵ satisfying

$$\mathbb{P}\left(F(x_\epsilon) - F^* \leq \epsilon\right) \geq \frac{2}{3}$$

at a cost $\mathcal{C}_{\mathcal{M}}(\epsilon, F)$.

A formal question

$$\min_x \ F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

Minimization Oracle: Suppose $\mathcal{M}_F(\epsilon)$ returns x_ϵ satisfying

$$\mathbb{P}\left(F(x_\epsilon) - F^* \leq \epsilon\right) \geq \frac{2}{3}$$

at a cost $\mathcal{C}_{\mathcal{M}}(\epsilon, F)$. [e.g. $\mathbb{E}[F(x_\epsilon) - F^*] \leq \epsilon/3$]

A formal question

$$\min_x \; F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

Minimization Oracle: Suppose $\mathcal{M}_F(\epsilon)$ returns x_ϵ satisfying

$$\mathbb{P}\left(F(x_\epsilon) - F^* \leq \epsilon\right) \geq \frac{2}{3}$$

at a cost $\mathcal{C}_{\mathcal{M}}(\epsilon, F)$. [e.g. $\mathbb{E}[F(x_\epsilon) - F^*] \leq \epsilon/3$]

Question:

Does there exist a procedure with high confidence guarantee

$$\mathbb{P}\left(F(x) - F^* \leq \epsilon\right) \geq 1 - p$$

and total cost $\sim \log(\frac{1}{p}) \cdot \mathcal{C}_{\mathcal{M}}(\epsilon, F)$?

A formal question

$$\min_x \; F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

Minimization Oracle: Suppose $\mathcal{M}_F(\epsilon)$ returns x_ϵ satisfying

$$\mathbb{P}\left(F(x_\epsilon) - F^* \le \epsilon\right) \ge \frac{2}{3}$$

at a cost $\mathcal{C}_{\mathcal{M}}(\epsilon, F)$. [e.g. $\mathbb{E}[F(x_\epsilon) - F^*] \le \epsilon/3$]

Question:

Does there exist a procedure with high confidence guarantee

$$\mathbb{P}\left(F(x) - F^* \le \epsilon\right) \ge 1 - p$$

and total cost $\sim \log(\frac{1}{p}) \cdot \mathcal{C}_{\mathcal{M}}(\epsilon, F)$?

Answer: Yes! for most interesting algorithms and

$$\text{cost} \quad \sim \quad \log(\kappa) \cdot \log\left(\tfrac{1}{p}\right) \cdot \mathcal{C}_{\mathcal{M}}\left(\frac{\epsilon}{\log(\kappa)}, F\right)$$

Naive attempt: sample and check

$$\min_{x} \; F(x) = \mathbb{E}_z[f(x,z)]$$

Strategy: generate $x_1, x_2, \ldots, x_m$ using $\mathcal{M}_f(\epsilon)$ and compute

$$\min_{i=1,\ldots,m} F(x_i).$$

Naive attempt: sample and check

$$\min_x \; F(x) = \mathbb{E}_z[f(x,z)]$$

Strategy: generate $x_1, x_2, \ldots, x_m$ using $\mathcal{M}_f(\epsilon)$ and compute

$$\min_{i=1,\ldots,m} F(x_i).$$

Sample efficiency: Evaluation is mean estimation, and typically costs

$$\Omega\left(\frac{\sigma^2 \ln(\frac{1}{p})}{\epsilon^2}\right) \quad \overset{\text{can be}}{\gg} \quad \mathcal{C}_{\mathcal{M}}(\epsilon, F).$$

Naive attempt: sample and check

$$\min_{x} \; F(x) = \mathbb{E}_z[f(x,z)]$$

Strategy: generate $x_1, x_2, \ldots, x_m$ using $\mathcal{M}_f(\epsilon)$ and compute

$$\min_{i=1,\ldots,m} F(x_i).$$

Sample efficiency: Evaluation is mean estimation, and typically costs

$$\Omega\left(\frac{\sigma^2 \ln(\frac{1}{p})}{\epsilon^2}\right) \quad \overset{\text{can be}}{\gg} \quad \mathcal{C}_{\mathcal{M}}(\epsilon, F).$$

Different approach: `proxBoost` uses two ingredients

1. robust distance estimation
   (Nemirovsky-Yudin '83, Minsker '16, Hsu-Sabato '16)
2. proximal point method
   (Moreau '65, Martinet '70, Rockafellar '76,. . .)

Robust distance estimation

Recall: smoothness+strong convexity

$$\mu \leq \frac{F(x) - \min F}{\frac{1}{2}\|x - \bar{x}\|^2} \leq L$$

Robust distance estimation

Recall: smoothness+strong convexity

$$\mu \le \frac{F(x) - \min F}{\frac{1}{2}\|x - \bar{x}\|^2} \le L$$

Idea:

$$F(x_\epsilon) - F^* \le \epsilon \qquad \Longrightarrow \qquad \|x_\epsilon - \bar{x}\| \le \sqrt{\frac{2\epsilon}{\mu}}$$

Robust distance estimation

Recall: smoothness+strong convexity

$$\mu \leq \frac{F(x) - \min F}{\frac{1}{2}\|x - \bar{x}\|^2} \leq L$$

Idea:

$$F(x_\epsilon) - F^* \leq \epsilon \qquad \Longrightarrow \qquad \|x_\epsilon - \bar{x}\| \leq \sqrt{\frac{2\epsilon}{\mu}} =: \delta.$$

Robust distance estimation

Recall: smoothness+strong convexity

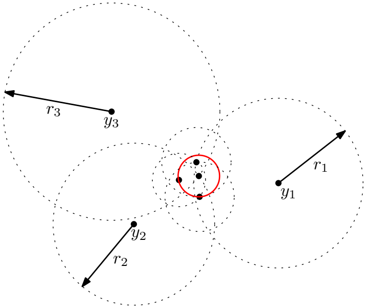
$$\mu \le \frac{F(x) - \min F}{\frac{1}{2}\|x - \bar{x}\|^2} \le L$$

Idea:

$$F(x_\epsilon) - F^* \le \epsilon \qquad \Longrightarrow \qquad \|x_\epsilon - \bar{x}\| \le \sqrt{\frac{2\epsilon}{\mu}} =: \delta.$$

Robust Distance Estimator (RDE):

- generate $\mathcal{Y} = \{y_1, \ldots, y_m\}$ by $\mathcal{M}_F(\epsilon)$.
- set $r_i = \min\{r \ge 0 : |B_r(y_i) \cap \mathcal{Y}| > \frac{m}{2}\}$.
- return y_{i^*} with $i^* = \arg\min_i r_i$.



Robust distance estimation

Recall: smoothness+strong convexity

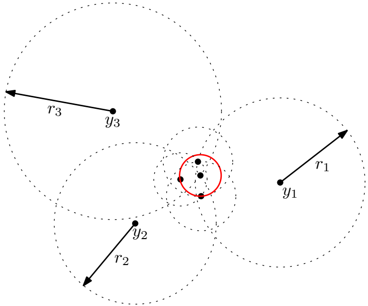
$$\mu \le \frac{F(x) - \min F}{\frac{1}{2}\|x - \bar{x}\|^2} \le L$$

Idea:

$$F(x_\epsilon) - F^* \le \epsilon \qquad \Longrightarrow \qquad \|x_\epsilon - \bar{x}\| \le \sqrt{\frac{2\epsilon}{\mu}} =: \delta.$$

Robust Distance Estimator (RDE):

- generate $\mathcal{Y} = \{y_1, \ldots, y_m\}$ by $\mathcal{M}_F(\epsilon)$.
- set $r_i = \min\{r \ge 0 : |B_r(y_i) \cap \mathcal{Y}| > \frac{m}{2}\}$.
- return y_{i^*} with $i^* = \arg\min_i r_i$.



Thm:(Nemirovsky-Yudin '83, Hsu-Sabato '16) Setting $m \approx \log\left(\frac{1}{p}\right)$ yields

$$\mathbb{P}\left(\|y_{i^*} - \bar{x}\| \lesssim \delta\right) \ge 1 - p$$

Therefore $F(y_{i^*}) - F^* \lesssim \kappa\epsilon$ with probability $1 - p$. (How to remove κ?)

proxBoost

Algorithm 1: `proxBoost`(δ, p, T)

Let x_{-1} be generated by **RDE** using $\mathcal{M}(\epsilon)$ on F.

Step $t = 0, \ldots, T$:

Let x_{t+1} be generated by **RDE** using $\mathcal{M}(\epsilon)$ on

$$F^{(t)}(x) := F(x) + \frac{\mu 2^t}{2} \|x - x_t\|^2$$

proxBoost

Algorithm 2: `proxBoost`(δ, p, T)

Let x_{-1} be generated by **RDE** using $\mathcal{M}(\epsilon)$ on F.

Step $t = 0, \ldots, T$:

Let x_{t+1} be generated by **RDE** using $\mathcal{M}(\epsilon)$ on

$$F^{(t)}(x) := F(x) + \frac{\mu 2^t}{2} \|x - x_t\|^2$$

Thm: (Davis-D-Xiao-Zhang '19) Setting $m \approx \log\left(\frac{1}{p}\right)$, $T \approx \log(\kappa)$ guarantees

$$\mathbb{P}\left(F(x_T) - F^* \lesssim \log(\kappa)\epsilon\right) \geq 1 - \log(\kappa)p.$$

Application 1: stochastic gradient

$$\min_x \; F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

Stochastic gradient.

$$\left\{ \begin{aligned} &\text{Sample } z_t \sim P \\ &\text{Set } x_{t+1} = x_t - \alpha_t \nabla f(x_t, z_t) \end{aligned} \right\}$$

Application 1: stochastic gradient

$$\min_x \; F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

Stochastic gradient.

$$\begin{cases} \text{Sample } z_t \sim P \\ \text{Set } x_{t+1} = x_t - \alpha_t \nabla f(x_t, z_t) \end{cases}$$

Sample complexity. Define condition number $\kappa = \frac{L}{\mu}$ and assume

$$\mathbb{E}_z \|\nabla f(x,z) - \nabla F(x)\|^2 \leq \sigma^2 \qquad \forall x.$$

Stochastic gradient	$\mathcal{O}\left(\kappa \cdot \ln\left(\frac{F(x_0)-F^*}{\epsilon}\right) + \frac{\sigma^2}{\mu\epsilon}\right)$
Accelerated SG	$\mathcal{O}\left(\sqrt{\kappa} \cdot \ln\left(\frac{F(x_0)-F^*}{\epsilon}\right) + \frac{\sigma^2}{\mu\epsilon}\right)$

Table: Samples for $\mathbb{E}[F(x) - F^*] \leq \epsilon$ (Ghadimi-Lan '13)

Application 1: stochastic gradient

$$\min_x \; F(x) = \mathbb{E}_{z \sim P}[f(x,z)]$$

Stochastic gradient.

$$\begin{Bmatrix} \text{Sample } z_t \sim P \\ \text{Set } x_{t+1} = x_t - \alpha_t \nabla f(x_t, z_t) \end{Bmatrix}$$

Sample complexity. Define condition number $\kappa = \frac{L}{\mu}$ and assume

$$\mathbb{E}_z \|\nabla f(x,z) - \nabla F(x)\|^2 \le \sigma^2 \qquad \forall x.$$

Stochastic gradient	$\mathcal{O}\left(\kappa \cdot \ln\left(\frac{F(x_0)-F^*}{\epsilon}\right) + \frac{\sigma^2}{\mu\epsilon}\right)$
Accelerated SG	$\mathcal{O}\left(\sqrt{\kappa} \cdot \ln\left(\frac{F(x_0)-F^*}{\epsilon}\right) + \frac{\sigma^2}{\mu\epsilon}\right)$

Table: Samples for $\mathbb{E}[F(x) - F^*] \le \epsilon$ (Ghadimi-Lan '13)

Remark: High confidence bounds available when

- sub-Gaussian gradients for SG & accelerated SG (Ghadimi-Lan '13)
- for truncated SG (Juditsky et al. '19)

Application 1: stochastic gradient

$$\min_x \; F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

Stochastic gradient.

$$\begin{Bmatrix} \text{Sample } z_t \sim P \\ \text{Set } x_{t+1} = x_t - \alpha_t \nabla f(x_t, z_t) \end{Bmatrix}$$

Sample complexity. Define condition number $\kappa = \frac{L}{\mu}$ and assume

$$\mathbb{E}_z \|\nabla f(x,z) - \nabla F(x)\|^2 \leq \sigma^2 \qquad \forall x.$$

Stochastic gradient	$\mathcal{O}\left(\kappa \cdot \ln\left(\frac{F(x_0)-F^*}{\epsilon}\right) + \frac{\sigma^2}{\mu\epsilon}\right)$
Accelerated SG	$\mathcal{O}\left(\sqrt{\kappa} \cdot \ln\left(\frac{F(x_0)-F^*}{\epsilon}\right) + \frac{\sigma^2}{\mu\epsilon}\right)$

Table: Samples for $\mathbb{E}[F(x) - F^*] \leq \epsilon$ (Ghadimi-Lan '13)

Remark: High confidence bounds available when

- sub-Gaussian gradients for SG & accelerated SG (Ghadimi-Lan '13)
- for truncated SG (Juditsky et al. '19)

(proxBoost+accelerated SGD): high confidence, overhead $\sim \log(1/p) \cdot \log(\kappa)$.

Application 1: stochastic gradient

$$\min_x \; F(x) = \mathbb{E}_{z \sim P}[f(x,z)]$$

Stochastic gradient.

$$\begin{cases} \text{Sample } z_t \sim P \\ \text{Set } x_{t+1} = x_t - \alpha_t \nabla f(x_t, z_t) \end{cases}$$

Sample complexity. Define condition number $\kappa = \frac{L}{\mu}$ and assume

$$\mathbb{E}_z \|\nabla f(x,z) - \nabla F(x)\|^2 \leq \sigma^2 \qquad \forall x.$$

Stochastic gradient	$\mathcal{O}\left(\kappa \cdot \ln\left(\frac{F(x_0)-F^*}{\epsilon}\right) + \frac{\sigma^2}{\mu\epsilon}\right)$
Accelerated SG	$\mathcal{O}\left(\sqrt{\kappa} \cdot \ln\left(\frac{F(x_0)-F^*}{\epsilon}\right) + \frac{\sigma^2}{\mu\epsilon}\right)$

Table: Samples for $\mathbb{E}[F(x) - F^*] \leq \epsilon$ (Ghadimi-Lan '13)

Remark: High confidence bounds available when

- sub-Gaussian gradients for SG & accelerated SG (Ghadimi-Lan '13)
- for truncated SG (Juditsky et al. '19)

(proxBoost+accelerated SGD): high confidence, overhead $\sim \log(1/p) \cdot \log(\kappa)$.

Gorbunov-Danilova-Gasnikov '20 (70 p): direct method, overhead $\sim \log(1/p)$

Example 2: empirical risk minimization

$$\min_x \; F(x) = \mathbb{E}_{z \sim P}[f(x, z)]$$

Empirical risk minimization. Sample $z_1, \ldots, z_n \sim P$ and solve

$$x_S = \arg\min_x \; F_S(x) = \frac{1}{n} \sum_{i=1}^{n} f(x, z_i)$$

Example 2: empirical risk minimization

$$\min_x \; F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

Empirical risk minimization. Sample $z_1, \ldots, z_n \sim P$ and solve

$$x_S = \arg\min_x \; F_S(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, z_i)$$

Sample complexity. Assume

- $f(\cdot, z)$ is nonnegative,
- F_S is μ-strongly convex w.p. $\frac{5}{6}$ when $n \geq N$.

Hsu-Sabato '16 propose an ERM-based estimator satisfying

$$\mathbb{P}\left(\frac{F(x) - F^*}{F^*} \leq \gamma\right) \geq 1 - p \qquad \text{whenever} \qquad n \gtrsim \ln\left(\tfrac{1}{p}\right) \max\left\{\frac{\kappa^2}{\gamma}, N\right\}$$

Example 2: empirical risk minimization

$$\min_x \; F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

Empirical risk minimization. Sample $z_1, \ldots, z_n \sim P$ and solve

$$x_S = \arg\min_x \; F_S(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, z_i)$$

Sample complexity. Assume

- $f(\cdot, z)$ is nonnegative,
- F_S is μ-strongly convex w.p. $\frac{5}{6}$ when $n \geq N$.

Hsu-Sabato '16 propose an ERM-based estimator satisfying

$$\mathbb{P}\left(\frac{F(x) - F^*}{F^*} \leq \gamma\right) \geq 1 - p \qquad \text{whenever} \qquad n \gtrsim \ln\left(\tfrac{1}{p}\right) \max\left\{\frac{\kappa^2}{\gamma}, N\right\}$$

(proxBoost+ERM): better sample efficiency $\sim \log(1/p) \cdot \log^2(\kappa) \cdot \max\left\{\frac{\kappa}{\gamma}, N\right\}$.

Extension: constraints and regularizers

Problem.

$$\min_{x \in \mathcal{X}} \; F(x) = \mathbb{E}_{z \sim P}[f(x, z)]$$

where $f(\cdot, z)$ as before and $\mathcal{X}$ is closed convex.

Extension: constraints and regularizers

Problem.

$$\min_{x \in \mathcal{X}} \; F(x) = \mathbb{E}_{z \sim P}[f(x,z)]$$

where $f(\cdot, z)$ as before and $\mathcal{X}$ is closed convex.

Back to basics: smoothness+strong convexity

$$\langle \nabla F(\bar{x}), x - \bar{x} \rangle + \frac{\mu}{2}\|x - \bar{x}\|^2 \le F(x) - \min_{\mathcal{X}} F \le \langle \nabla F(\bar{x}), x - \bar{x} \rangle + \frac{L}{2}\|x - \bar{x}\|^2$$

Extension: constraints and regularizers

Problem.

$$\min_{x \in \mathcal{X}} \; F(x) = \mathbb{E}_{z \sim P}[f(x, z)]$$

where $f(\cdot, z)$ as before and $\mathcal{X}$ is closed convex.

Back to basics: smoothness+strong convexity

$$\langle \nabla F(\bar{x}), x - \bar{x} \rangle + \frac{\mu}{2}\|x - \bar{x}\|^2 \leq F(x) - \min_{\mathcal{X}} F \leq \langle \nabla F(\bar{x}), x - \bar{x} \rangle + \frac{L}{2}\|x - \bar{x}\|^2$$

Idea:

$$F(x_\epsilon) - \min_{\mathcal{X}} F \leq \epsilon \qquad \Longrightarrow \qquad \left\{ \begin{array}{c} \|x_\epsilon - \bar{x}\| \leq \sqrt{\dfrac{2\epsilon}{\mu}} \\ 0 \leq \langle \nabla F(\bar{x}), x_\epsilon - \bar{x} \rangle \leq \epsilon \end{array} \right\}$$

Extension: constraints and regularizers

Problem.

$$\min_{x\in\mathcal{X}} \; F(x) = \mathbb{E}_{z\sim P}[f(x,z)]$$

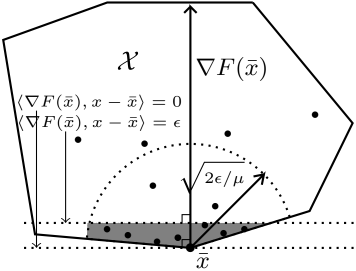
where $f(\cdot, z)$ as before and $\mathcal{X}$ is closed convex.

Back to basics: smoothness+strong convexity

$$\langle \nabla F(\bar{x}), x-\bar{x}\rangle + \frac{\mu}{2}\|x-\bar{x}\|^2 \le F(x) - \min_{\mathcal{X}} F \le \langle \nabla F(\bar{x}), x-\bar{x}\rangle + \frac{L}{2}\|x-\bar{x}\|^2$$

Idea:

$$F(x_\epsilon) - \min_{\mathcal{X}} F \le \epsilon \quad\Longrightarrow\quad \left\{ \begin{array}{c} \|x_\epsilon - \bar{x}\| \le \sqrt{\dfrac{2\epsilon}{\mu}} \\ 0 \le \langle \nabla F(\bar{x}), x_\epsilon - \bar{x}\rangle \le \epsilon \end{array} \right\}$$



A sketch...

Conceptual algorithm: Apply RDE in the metric

$$\rho(x, x') = \max\left\{\sqrt{\frac{\epsilon\mu}{2}} \cdot \|x - x'\|, |\langle \nabla F(\bar{x}), x - x' \rangle|\right\}.$$

A sketch...

Conceptual algorithm: Apply RDE in the metric

$$\rho(x,x') = \max\left\{\sqrt{\frac{\epsilon\mu}{2}}\cdot\|x-x'\|, |\langle\nabla F(\bar{x}), x-x'\rangle|\right\}.$$

After $m \approx \log(\frac{1}{p})$ rounds, get

$$\mathbb{P}(F(x) - F(\bar{x}) \lesssim \kappa\epsilon) \geq 1-p.$$

Use `proxBoost` to remove κ.

A sketch...

Conceptual algorithm: Apply RDE in the metric

$$\rho(x, x') = \max\left\{\sqrt{\frac{\epsilon\mu}{2}} \cdot \|x - x'\|, |\langle \nabla F(\bar{x}), x - x' \rangle|\right\}.$$

After $m \approx \log(\frac{1}{p})$ rounds, get

$$\mathbb{P}(F(x) - F(\bar{x}) \lesssim \kappa\epsilon) \geq 1 - p.$$

Use `proxBoost` to remove κ.

But... we don't know $\nabla F(\bar{x})$.

A sketch...

Conceptual algorithm: Apply RDE in the metric

$$\rho(x,x') = \max\left\{\sqrt{\frac{\epsilon\mu}{2}}\cdot\|x-x'\|, |\langle\nabla F(\bar{x}), x-x'\rangle|\right\}.$$

After $m \approx \log(\frac{1}{p})$ rounds, get

$$\mathbb{P}(F(x) - F(\bar{x}) \lesssim \kappa\epsilon) \geq 1 - p.$$

Use `proxBoost` to remove κ.

But... we don't know $\nabla F(\bar{x})$.

Pre-processing: Replace $\nabla F(\bar{x})$ with estimator $\hat{\nabla} F(\hat{x})$ where

(i) $\hat{x}$ generated using Euclidean RDE as before,

(ii) $$\hat{\nabla} F(\hat{x}) := \frac{1}{n}\sum_{i=1}^{n} \nabla f(\hat{x}, z_i)$$

A sketch...

Conceptual algorithm: Apply RDE in the metric

$$\rho(x, x') = \max\left\{\sqrt{\frac{\epsilon\mu}{2}} \cdot \|x - x'\|, |\langle \nabla F(\bar{x}), x - x' \rangle|\right\}.$$

After $m \approx \log(\frac{1}{p})$ rounds, get

$$\mathbb{P}(F(x) - F(\bar{x}) \lesssim \kappa\epsilon) \geq 1 - p.$$

Use `proxBoost` to remove κ.

But... we don't know $\nabla F(\bar{x})$.

Pre-processing: Replace $\nabla F(\bar{x})$ with estimator $\hat{\nabla} F(\hat{x})$ where

(i) $\hat{x}$ generated using Euclidean RDE as before,

(ii) $$\hat{\nabla} F(\hat{x}) := \frac{1}{n}\sum_{i=1}^{n} \nabla f(\hat{x}, z_i) \qquad \text{with} \qquad n \approx \frac{1}{\kappa^2} \cdot \frac{\mathrm{Var}(\nabla f(\hat{x}, \cdot))}{\mu\epsilon}$$

A sketch...

Conceptual algorithm: Apply RDE in the metric

$$\rho(x, x') = \max\left\{\sqrt{\frac{\epsilon\mu}{2}} \cdot \|x - x'\|, |\langle \nabla F(\bar{x}), x - x'\rangle|\right\}.$$

After $m \approx \log(\frac{1}{p})$ rounds, get

$$\mathbb{P}(F(x) - F(\bar{x}) \lesssim \kappa\epsilon) \geq 1 - p.$$

Use `proxBoost` to remove κ.

But... we don't know $\nabla F(\bar{x})$.

Pre-processing: Replace $\nabla F(\bar{x})$ with estimator $\hat{\nabla} F(\hat{x})$ where

(i) $\hat{x}$ generated using Euclidean RDE as before,

(ii) $$\hat{\nabla} F(\hat{x}) := \frac{1}{n} \sum_{i=1}^{n} \nabla f(\hat{x}, z_i) \qquad \text{with} \qquad n \approx \frac{1}{\kappa^2} \cdot \frac{\mathrm{Var}(\nabla f(\hat{x}, \cdot))}{\mu\epsilon}$$

Why does this work?

- Since $\|x_i - \bar{x}\| \approx \sqrt{\frac{\epsilon}{\mu}}$, it suffices to ensure $\|\nabla F(\bar{x}) - \hat{\nabla} F(\hat{x})\| \lesssim \kappa\sqrt{\mu\epsilon}$.

Thank you!