OPTIMAL METHODS FOR RISK AVERSE OPTIMIZATION OVER A NETWORK

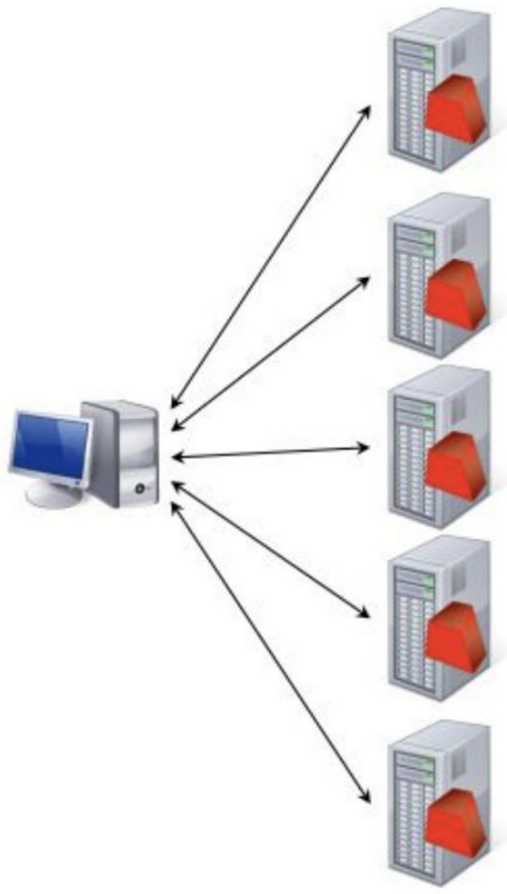
May 20, 2022

Guanghui (George) Lan and Zhe (Jimmy) Zhang,
ISyE, Georgia Tech

CREATING THE NEXT®

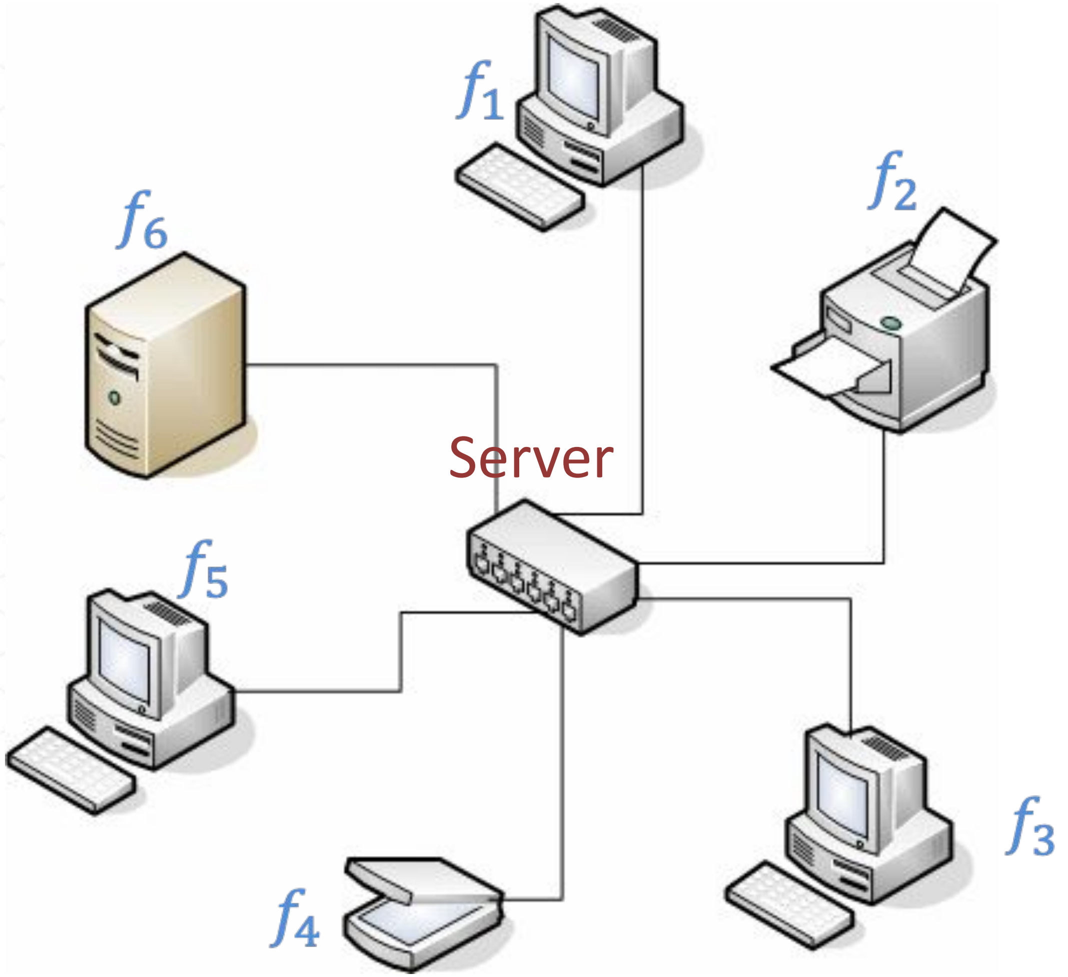
BEYOND RISK NEUTRAL?

- Risk Neutral Optimization
  - $min_{x \in X} \sum_{i=1}^{m} \frac{1}{m} f_i(x) + \mathrm{u}(x)$
- What if
  - One-time decision, e.g. Mars Landing Site
  - Downside risk e.g. financial portfolio
  - Empirical probability no good



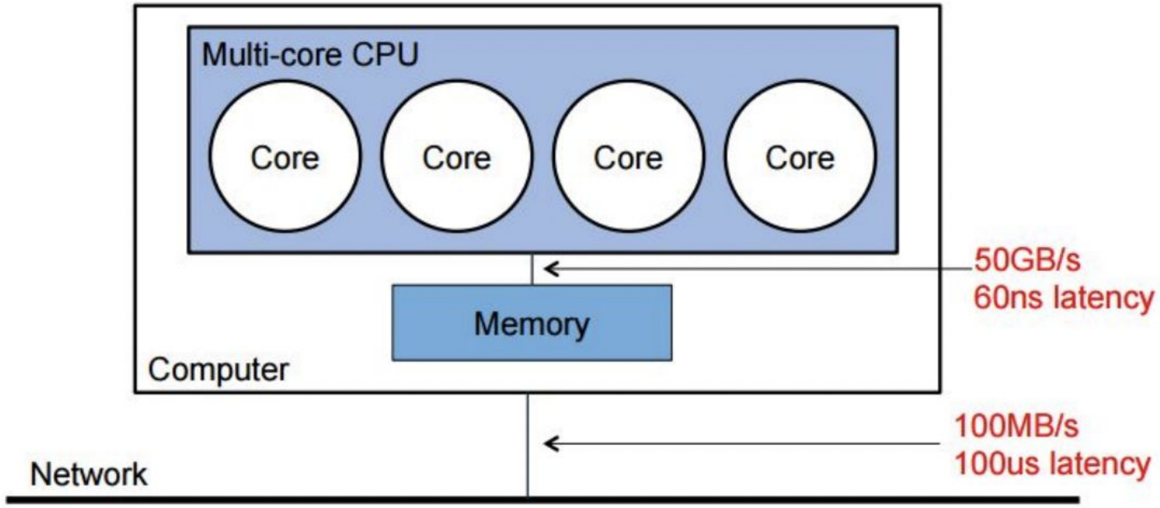
RISK-AVERSE OPTIMIZATION

- Risk Neutral
  - $min_{x \in X} \sum_{i=1}^{m} \bar{p_i} f_i(x) + \mathrm{u}(x)$
- Coherent Risk Measure ρ
  - $\min_{x \in X} \rho \, [f_1(x), f_2(x), \ldots, f_m(x)]$
  - $\min_{x \in X} \max_{p \in P} \sum_{i=1}^{m} p_i f_i(x) + \mathrm{u}(x) \; - \rho^*(p)$
  - Types of ρ :
    - CV@R
    - Mean Semideviation of order r, Entropic Risk
    - DRO ambiguity set



DISTRIBUTED OPTIMIZATION

- Communication is expensive:
  - L2 Cache Latency: 7ns
  - RAM ~ 60ns
  - Inside a cluster ~ 100us
  - LTE ~ 100ms

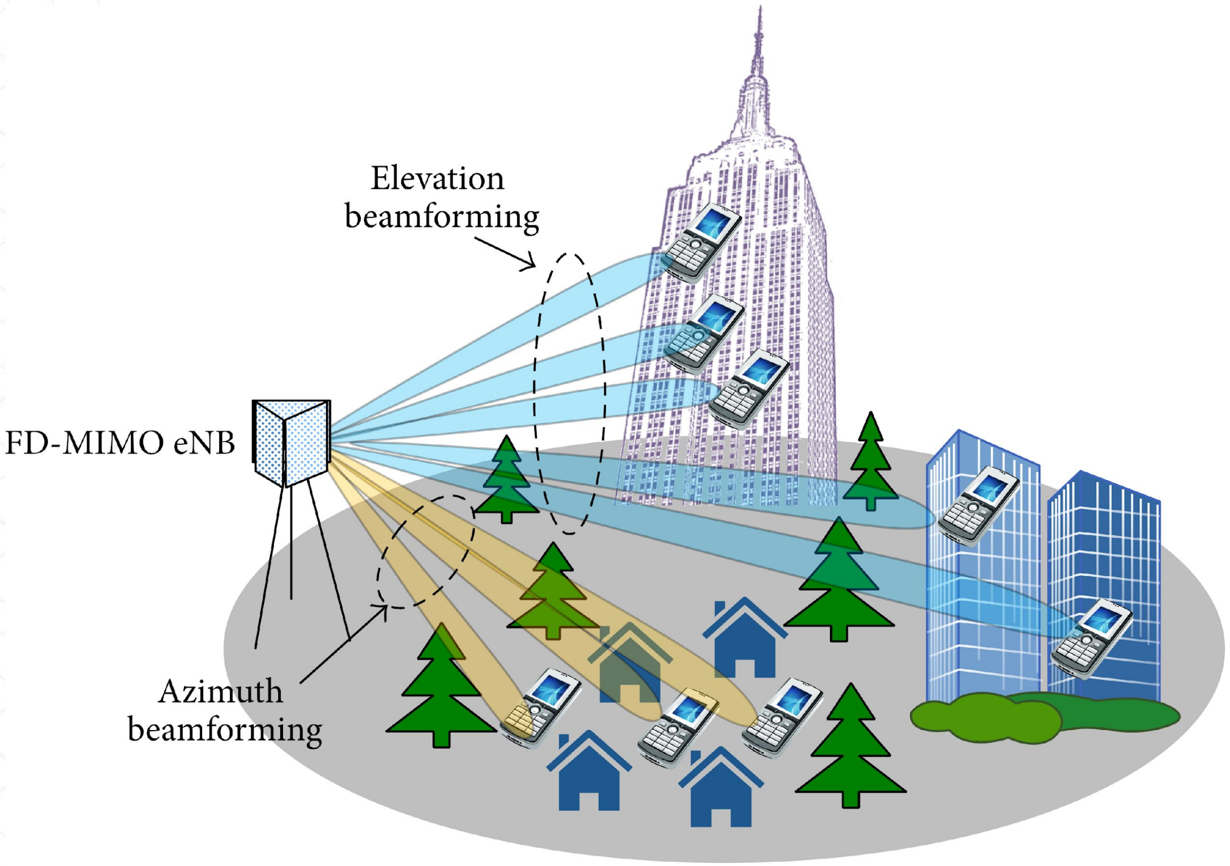


CLIMATE CHANGE PLANNING

- Infrastructure Investment for climate change mitigation
- ρ: CV@R corresponding 99% of possible scenarios
- $f_i(x)$: Long term economic cost under j^{th} climate model and k^{th} impact model.
  - Stored at the i^{th} (worker) computing node
- Few communication rounds $\Rightarrow$ Fast Computation

MIMO SYSTEM IN 5G COMMUNICATION

- Configure active antenna optimally ⇒ consistent speed for most users
- ρ: mean semi-deviation risk measure
- f_i: the negative downlink (uplink) speed
- Few exchange between terminal device and base station ⇒ more responsive base station



OUTLINE

$$\min_{x \in X} \rho\left[f_1(x), f_2(x), \dots, f_m(x)\right] + u(x)$$

Q: the least number of communication rounds for an ϵ-optimal solution? Can we solve it as easily as the risk-neutral problem?

- Communication-Efficient DRAO Method
- Communication and Computationally-Efficient DRAO-S Method
- Lower Communication Complexity Bound

DRAO: CHALLNEGE

- Consider smooth f_i's

$$min_{x \in X} \sum_{i=1}^{m} \frac{1}{m} f_i(x) + \mathrm{u}(x)$$

$$\Updownarrow$$

$$\min_{x \in X} \max_{p \in P} \sum_{i=1}^{m} p_i f_i(x) + \mathrm{u}(x) \ - \rho^*(p) + u(x)$$

DRAO: STRONG ORACLE

- We found from Nesterov (1998) that max-type function is essentially smoth

$$\max\{f_1(x), \ldots, f_m(x)\} \\ \leq \max\{f_1(\bar{x}) + \langle \nabla f_1(\bar{x}), x - \bar{x}\rangle, \ldots, f_m(\bar{x}) + \langle \nabla f_m(\bar{x}), x - \bar{x}\rangle\} + L_f \|x - \bar{x}\|^2 /2$$

- Prox-max-update

$$x^t \leftarrow \mathrm{argmin}_{x \in X} \max\left(f_1(\underline{x}) + \langle \nabla f_1(\underline{x}), x\rangle, \ldots, f_m(\underline{x}) + \langle \nabla f_m(\underline{x}), x\rangle\right) + \frac{\eta}{2}||x - \bar{x}||^2$$

  - Nesterov (1998) Nesterov Accelerated Gradient method, Lan (2015) Accelerated Prox-Level method

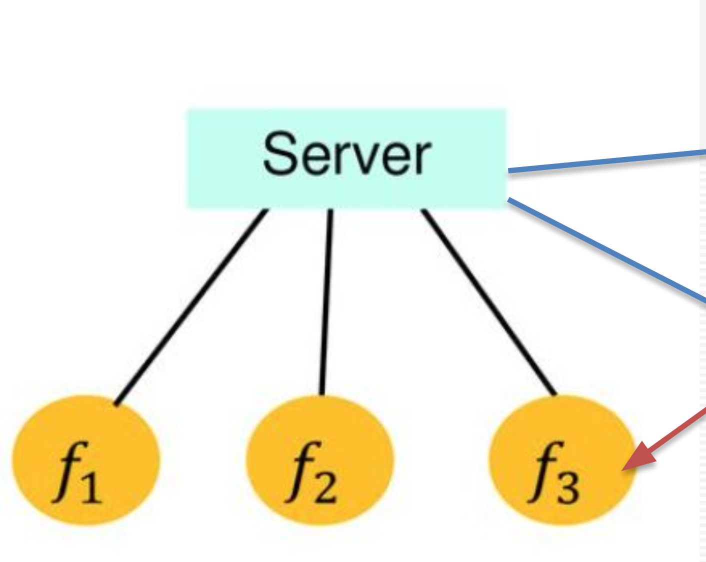
- Can we extend it
  - Coherent risk measure ρ, structured non-smooth function

DRAO: PRIMAL-DUAL TYPE METHOD

$$\min_{x \in X} \rho\,[f_1(x), f_2(x), \ldots, f_m(x)] + u(x)$$

Fenchel Linearization

$$\min_{x\in X} \max_{\pi\in\Pi} \rho\{\langle A_1 x, \pi_1\rangle - f_1^*(\pi_1), \ldots, \langle A_m x, \pi_m\rangle - f_m^*(\pi_m)\} + u(x)$$



Algorithm 1 A Generic **D**istributed **R**isk **A**verse **O**ptimization (DRAO) Method

1: $\tilde{x}^t \leftarrow x^{t-1} + \theta_t(x^{t-1} - x^{t-2})$.

2: $\pi_i^t \leftarrow \arg\max_{\pi_i\in\Pi_i} \langle A_i\tilde{x}^t, \pi_i\rangle - f_i^*(\pi_i) - \tau_t V_i(\pi_i; \pi_i^{t-1})$, and evaluates $v_i^t \leftarrow A_i^\top \pi_i^t$ and $f_i^*(\pi_i^t)$.

3: $x^t \leftarrow \arg\min_{x\in X} \rho\{\langle x, v_1^t\rangle - f_1^*(\pi_1^t), \ldots, \langle x, v_m^t\rangle - f_m^*(\pi_m^t)\} + u(x) + \frac{\eta_t}{2}\left\|x - x^{t-1}\right\|^2$.

DRAO: SMOOTH PROBLEM

- π_i-prox update on the worker

$$\pi_i^t \leftarrow \arg\max_{\pi_i \in \Pi_i} \langle A_i \tilde{x}^t, \pi_i \rangle - f_i^*(\pi_i) - \tau_t V_i(\pi_i; \pi_i^{t-1})$$

$\Leftrightarrow$

$$\underline{x}^t \leftarrow (\tilde{x}^t + \tau_t \underline{x}^{t-1})/(1+\tau_t),$$
$$\pi_i^t \leftarrow \nabla f_i(\underline{x}^t),$$
$$f_i^*(\pi_i^t) \leftarrow \langle \underline{x}^t, \pi_i^t \rangle - f_i(\underline{x}^t).$$

- Communication Complexity

$L_f := \max_{p \in P} L_{f,p}$, where $L_{f,p}$ is the smoothness cst for $\sum_i p_i f_i(x)$

	Convex ($\alpha = 0$)	strongly convex ($\alpha > 0$)
Smooth	$\mathcal{O}(\sqrt{L_f} R_0/\sqrt{\epsilon})$	$\mathcal{O}(\sqrt{L_f/\alpha} \log(1/\sqrt{\epsilon}))$

DRAO: STRUCTURED NONSMOOTH PROBLEM

- π_i-prox update on the worker

$$\pi_i^t \leftarrow \arg\max_{\pi_i \in \Pi_i} \langle A_i \tilde{x}^t, \pi_i \rangle - f_i^*(\pi_i) - \tau_t V_i(\pi_i; \pi_i^{t-1}) \quad \Longleftrightarrow \quad \pi_i^t \leftarrow \arg\max_{\pi_i \in \Pi_i} \langle A_i \tilde{x}^t, \pi_i \rangle - f_i^*(\pi_i) - \tfrac{\tau_t}{2} \left\|\pi_i - \pi_i^{t-1}\right\|^2 .$$

- Communication Complexity

$$M_A := \max_{p \in P} [\textstyle\sum_{i=1}^m p_i \|A_i\|_{2,2}^2]^{1/2}, \;\; D_\Pi := \max_{p \in P} [\max_{\pi, \bar{\pi} \in \Pi} \textstyle\sum_{i=1}^m p_i \|\pi_i - \bar{\pi}_i\|^2]^{1/2}$$

	Convex ($\alpha = 0$)	strongly convex ($\alpha > 0$)
Structured Non-smooth	$\mathcal{O}(M_A D_\Pi R_0/\epsilon)$	$\mathcal{O}(M_A D_\Pi/\sqrt{\epsilon\alpha})$

DRAO-S: COMPUTATION CHALLENGE

Georgia Tech

$$x^t \leftarrow \arg\min_{x\in X} \rho\{\langle x, v_1^t\rangle - f_1^*(\pi_1^t), \dots, \langle x, v_m^t\rangle - f_m^*(\pi_m^t)\} + u(x) + \tfrac{\eta_t}{2}\left\|x - x^{t-1}\right\|^2$$

- Hard risk measure ρ such that exact evaluation of prox-ρ-update is challenging
  - Mean upper-semi-deviation risk measure of order 2?
  - Kantorovich Ball?
- Access to P-prox oracle only:

$$\min_{x\in X} \max_{\pi\in\Pi} \rho\{\langle A_1 x, \pi_1\rangle - f_1^*(\pi_1), \dots, \langle A_m x, \pi_m\rangle - f_m^*(\pi_m)\} + u(x)$$

Fenchel Conjugate Again

$$\min_{x\in X} \max_{p\in P} \max_{\pi\in\Pi} \textstyle\sum_{i=1}^m p_i[\langle \pi_i, Ax\rangle - f_i^*(\pi_i)] - \rho^*(p) + u(x)$$

Q: Can we use only $O\left(\frac{1}{\epsilon}\right)$ P-projections while maintaining the same communication complexity?

SLIDING: A DELICATE TECHNIQUE

- Gradient sliding for additive composite function
  - Lan (2015) composite optimization

$$\min_{x\in X} f(x) + \boxed{g(x)}$$

Prox$_g$ easy

  - Lan (2021) graph topology invariant decentralized optimization

$$\min_{x\in X}\max_{\pi\in\Pi}\boxed{\max_{\lambda\geq 0}\langle x,\, L\lambda\rangle} + \langle \pi,\, x\rangle - g^*(\pi)$$

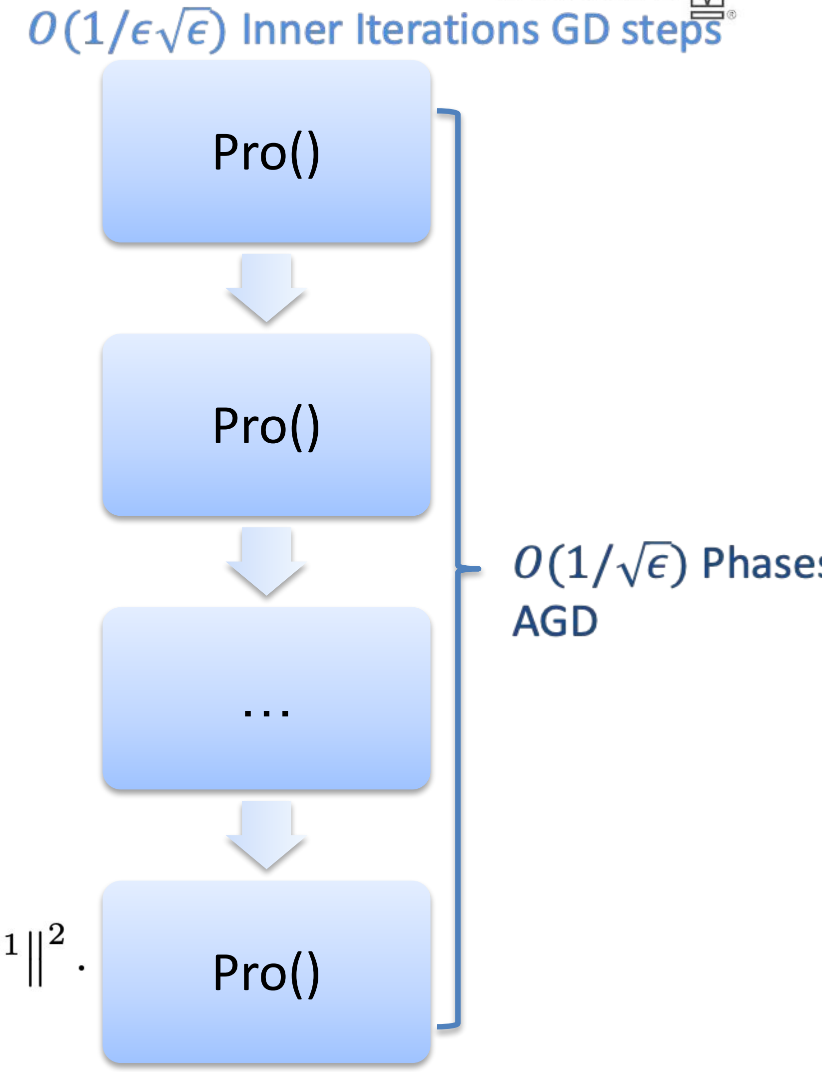
Prox$_\lambda$ easy

- Nested composite function

$$\min_{x\in X}\max_{\pi_i\in\Pi_i}\boxed{\max_{p\in P}\textstyle\sum_{i=1}^m p_i}[\langle \pi_i, A_i x\rangle - f_i^*(\pi_i)] + u(x)$$

Prox-ρ easy

$$x^t \leftarrow \arg\min_{x\in X}\max_{p\in P}\textstyle\sum_{i=1}^m p_i(\langle x, v_i^t\rangle - f_i^*(\pi_i^t)) - \rho^*(p) + u(x) + \frac{\eta_t}{2}\left\|x - x^{t-1}\right\|^2.$$



DRAO-S: SPS SUBROUTINE ON SERVER

$$x^t \leftarrow \arg\min_{x\in X}\max_{p\in P} \textstyle\sum_{i=1}^m p_i(\langle x, v_i^t\rangle - f_i^*(\pi_i^t)) - \rho^*(p) + u(x) + \frac{\eta_t}{2}\left\|x - x^{t-1}\right\|^2.$$

Algorithm 2 Saddle Point Sliding (SPS) Subroutine

Input: Initial points $x^{t-1}, y^0 \in X$, $p^0, p^{-1} \in P$, and gradients $\{v_i^t\}, \{v_i^{t-1}\}$. Non-negative stepsizes η_t, $\{\delta_s\}$, $\{\gamma_s\}$ and $\{\beta_s\}$, averaging weights $\{q_s\}$, and iteration number S_t.

1: **for** $s = 1, 2, 3 \ldots S_t$ **do**

2: $\tilde{v}^s \leftarrow \begin{cases} \sum_{i=1}^m p_i^0 v_i^t + \delta_1 \sum_{i=1}^m (p_i^0 - p_i^{-1}) v_i^{t-1} & \text{if } s = 1, \\ \sum_{i=1}^m p_i^{s-1} v_i^t + \delta_s \sum_{i=1}^m (p_i^{s-1} - p_i^{s-2}) v_i^t & \text{if } s \geq 2. \end{cases}$

3: $y^s \leftarrow \arg\min_{y\in X} \langle y, \tilde{v}^s\rangle + u(y) + \frac{\beta_s}{2}\left\|y - y^{s-1}\right\|^2 + \frac{\eta_t}{2}\left\|y - x^{t-1}\right\|^2.$

4: $p^s \leftarrow \arg\max_{p\in P} \sum_{i=1}^m p_i(\langle v_i^t, y^s\rangle - f_i^*(\pi_i^t)) - \rho^*(p) - \gamma_s U(p; p^{s-1}).$

5: **end for**

6: **return** $x^t := \sum_{s=1}^{S_t} q_s y^s / (\sum_{s=1}^{S_t} q_s)$, $y^t := y^{S_t}$, $\bar{p}^t := \sum_{s=1}^{S_t} q_s p^s / (\sum_{s=1}^{S_t} q_s)$, $p^t := p^{S_t}$ and $\tilde{p}^t = p^{S_t - 1}$.

DRAO-S: SMOOTH PROBLEM

- x^t in DRAO is generated instead by

$$(x^t, y^t, \bar{p}^t, p^t, \tilde{p}^t) = SPS(x^{t-1}, y^{t-1}, p^{t-1}, \tilde{p}^{t-1}, \{v_i^t\}, \{v_i^{t-1}\} \\ \mid \eta_t, \{\delta_s^t\}, \{\gamma_s^t\}, \{\beta_s^t\}, \{q_s^t\}, S_t) \ .$$

- Smooth Problem

$$M_t := ||v^t||_{2,U^*} := \max_{||p||_U \le 1, ||y|| \le 1} \sum_{i=1}^m p_i (v_i^t)^\top y$$

  - $S_t = \lceil t\, M_t\, \Delta \rceil \;\Rightarrow\; O\big(D_P \widetilde{M} R_0/\epsilon\big)$ P-projection oracle complexity
  - $\alpha > 0$: $S_t = \lceil (2\Delta/\theta^{t-1})^{1/2} \mathcal{M}_t \rceil \Rightarrow O(\kappa^{1/4} \tilde{M} D_P/(\alpha\sqrt{\epsilon}))$ P-projection oracle complexity

DRAO-S: STRUCTURED NON-SMOOTH

- Structured non-smooth problem

$$\tilde{M}_{A\Pi} = \max_{\pi\in\Pi}\{\|[A_1^\top \pi_1^t; \ldots; A_m^\top \pi_m^t]\|_{2,U^*} := \max_{\pi\in\Pi} \max_{\|y\|_2\le 1, \|p\|_U \le 1} \textstyle\sum_{i=1}^m p_i \langle A_i^\top \pi_i, y\rangle\}.$$

- Non-strongly convex problem
  - $S_t = \lceil M_t \ \Delta \rceil \Rightarrow$ $\mathcal{O}(\tilde{M}_{A\Pi} D_P R_0/\epsilon)$ P-projection oracle complexity

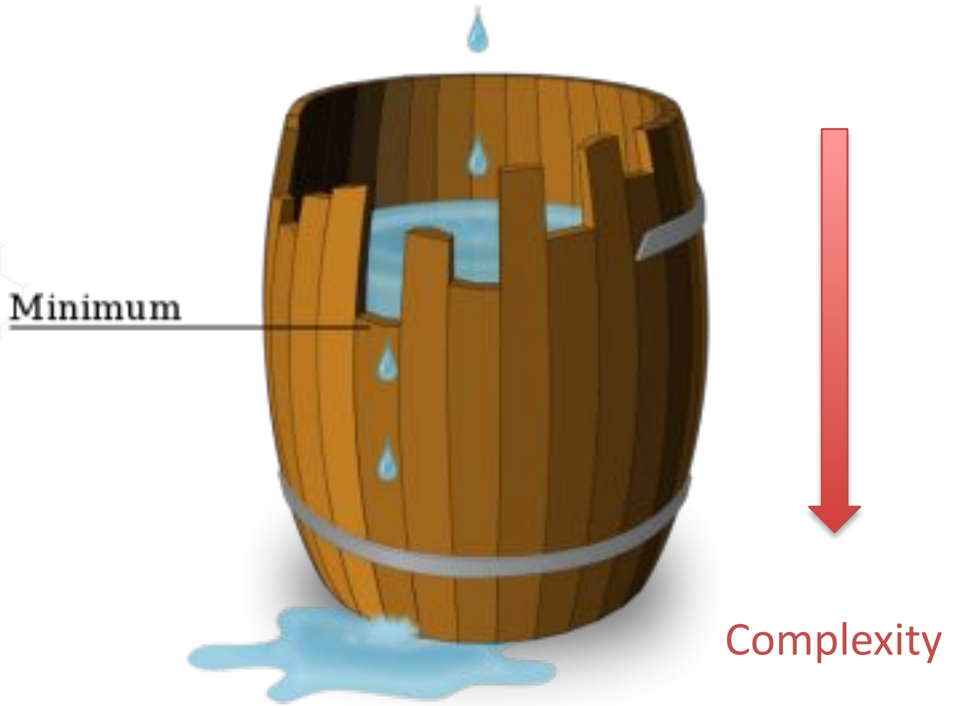
V.S. $\mathcal{O}(M_A D_\Pi R_0/\epsilon)$

- Strongly convex problem
  - $S_t = \lceil \tilde{M}_{A\Pi}^2 \Delta \rceil \Rightarrow \mathcal{O}(\tilde{M}_{A\Pi} D_P/\sqrt{\epsilon\alpha})$ P-projection oracle complexity

DRAO-S: TAKE-AWAY MESSAGE

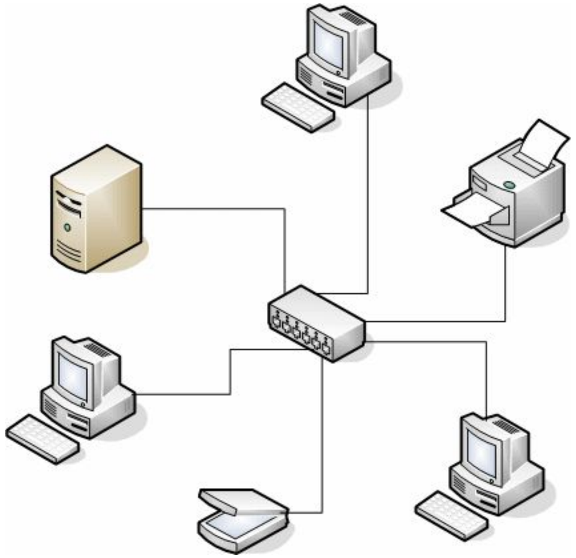
- Sliding is also possible for the nested composition

- In optimization, the individual complexity of a component in a problem is not limited by the complexity of the whole system.



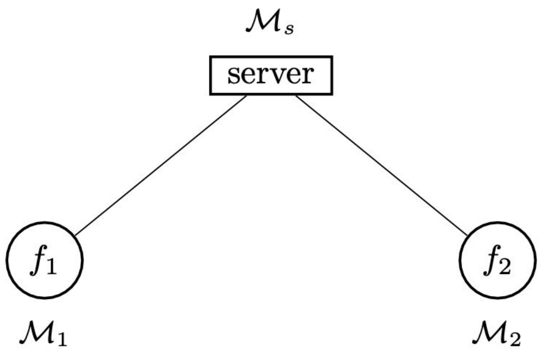
Q: *What's the **least** number of communication rounds to find an ϵ-optimal solution ?*

- Risk Neutral (Bach 17) $\mathcal{O}(\sqrt{L_{f,\bar{p}}}R_0/\sqrt{\epsilon})$ vs $\mathcal{O}(\sqrt{L_f}R_0/\sqrt{\epsilon})$
  - $L_{f,\bar{p}}$: Lipschitz smoothness constant of $\sum_{i=1}^n \frac{1}{n} f_i(x)$
  - L_f: Largest Lipschitz smoothness constant of among $\{\sum_{i=1}^n p_i f_i(x)\}_{p\in P}$
- Structured Non-smooth? More computation locally?



LOWER: COMPUTATION MODEL

- **Local Computation:** FO update (for arbitrary number of steps) of local memory, e.g. prox-update
- **Local memory:** all "reachable" points, linear span of evaluated gradients
- **Communication:** send anything from its memory
- **P Computation:** $\boldsymbol{p}$ is a linear combination weight. So automatically covered in the linear span framework.

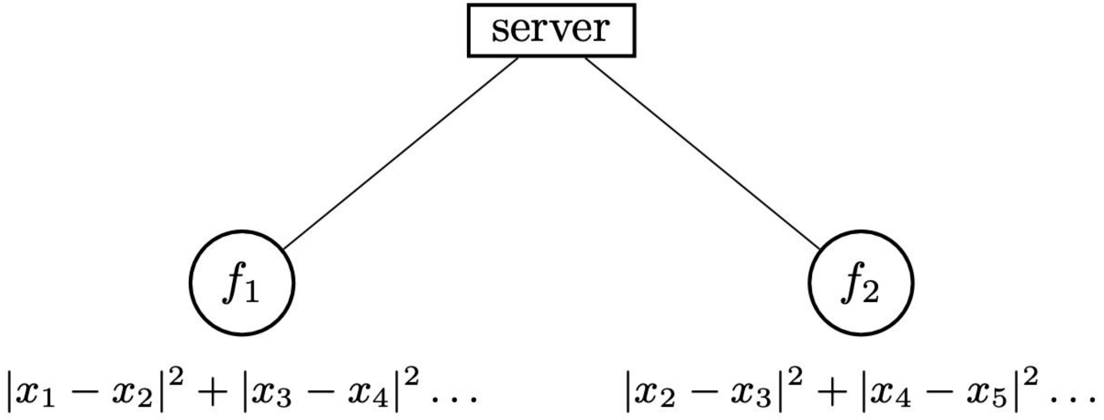


HARD INSTANCE

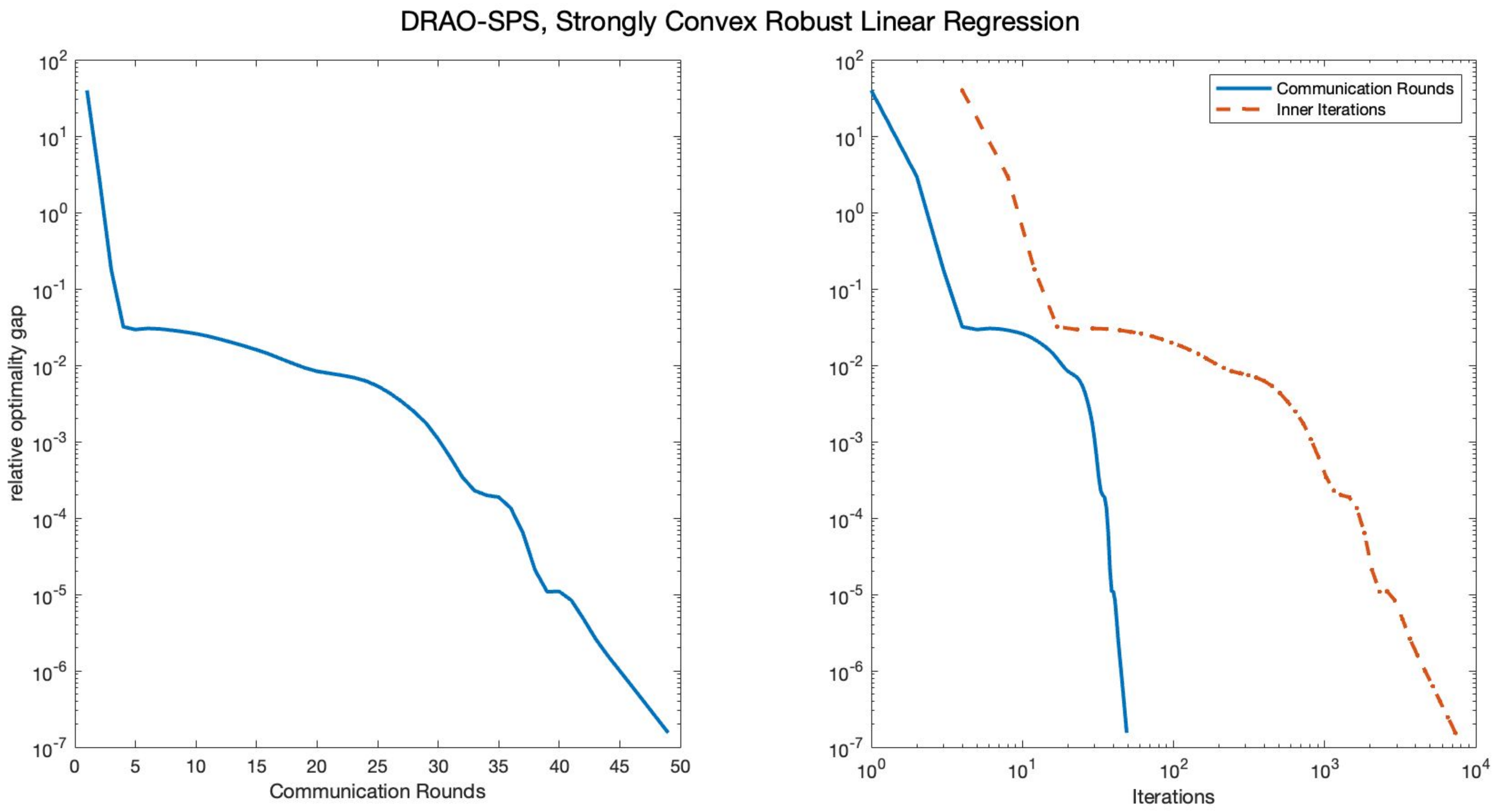
- Optimal is $x_i^* := (1 - \frac{i}{2k+2})\ \forall i \in [2k+1]$

- Structure

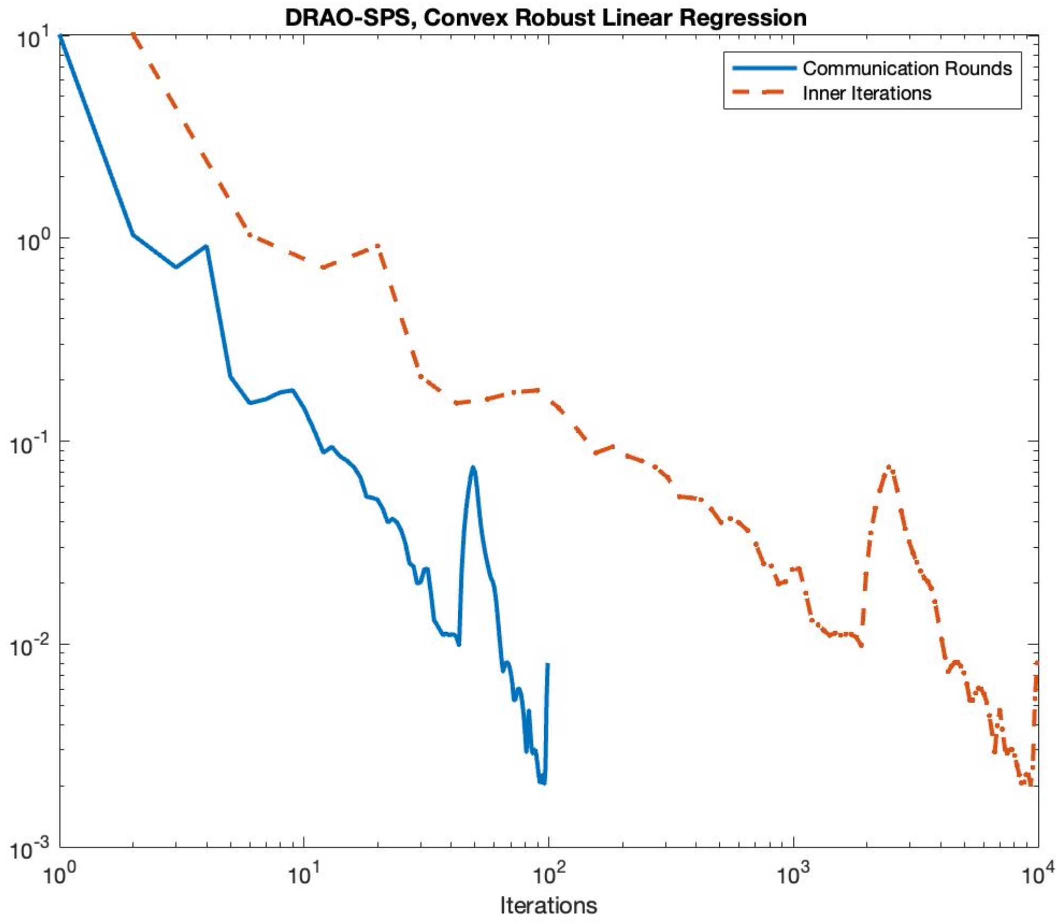
	Convex	Strongly convex
Smooth	$\mathcal{O}(\sqrt{L_f} R_0/\sqrt{\epsilon})$	$\mathcal{O}(\sqrt{L_f/\alpha} \log(1/\sqrt{\epsilon}))$
Structured Nonsmooth	$\mathcal{O}(M_A D_\Pi R_0/\epsilon)$	$\mathcal{O}(M_A D_\Pi/\sqrt{\epsilon\alpha})$



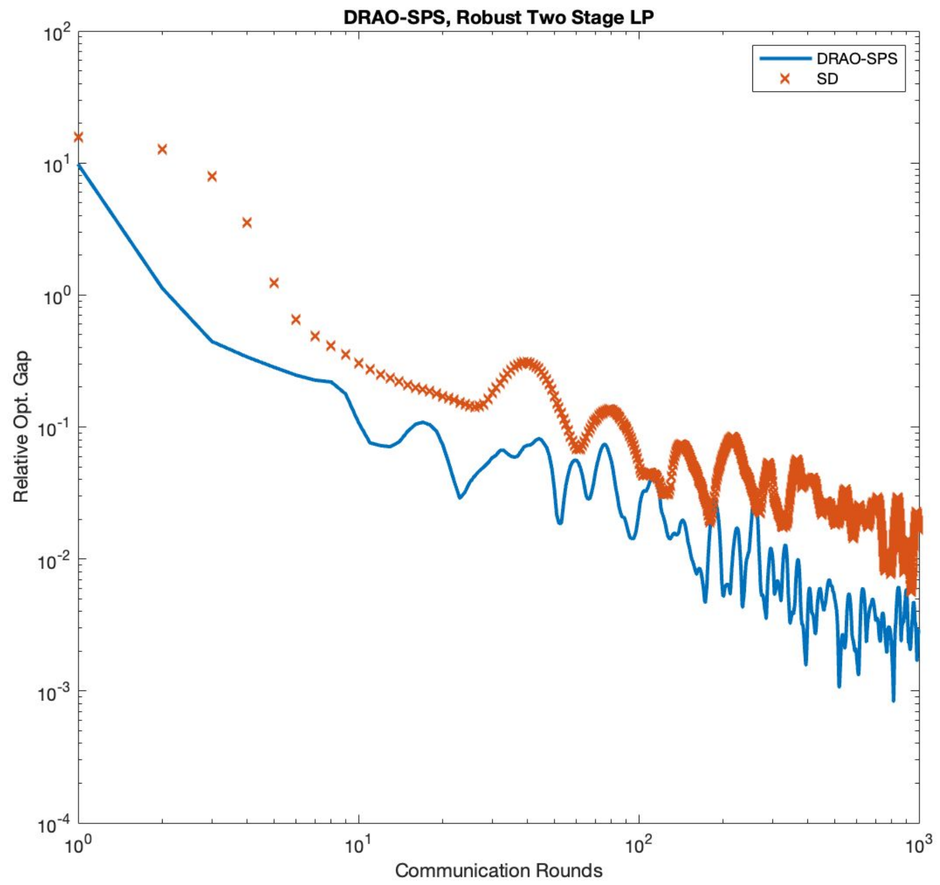
NUMERICAL : SMOOTH+STRONGLY CONVEX

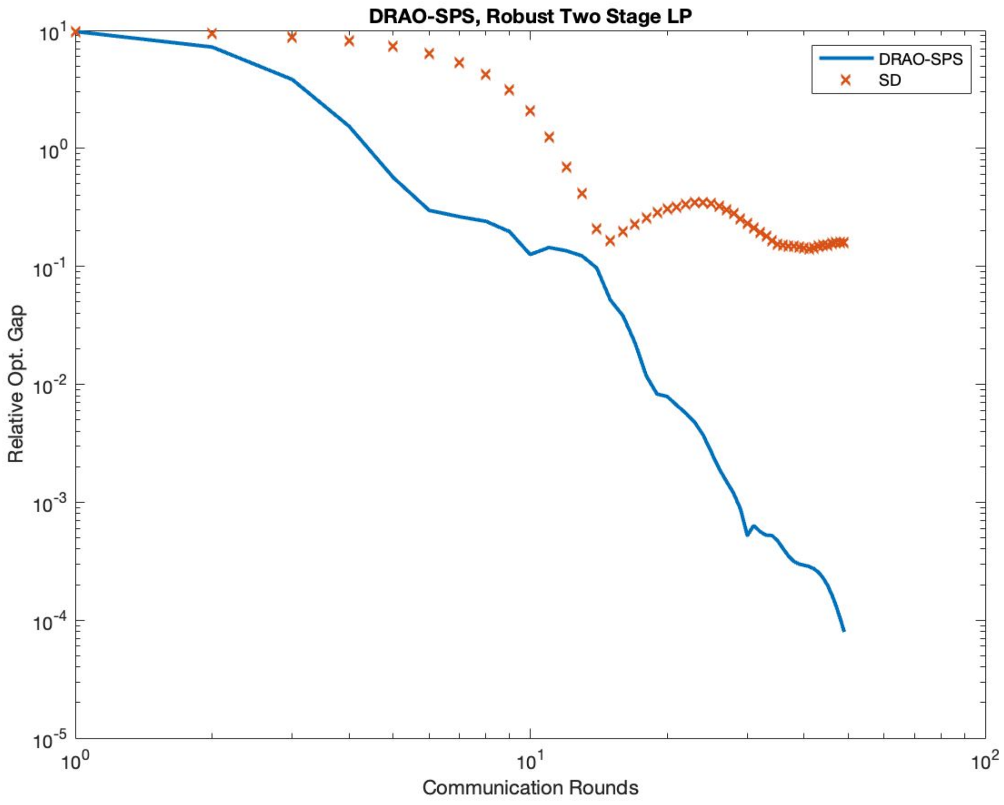


NUMERICAL: SMOOTH



NUMERICAL: STRUCTURED NON-SMOOTH V.S. SD METHOD





THANK YOU QUESTIONS?

Georgia Tech

- Risk Averse Optimization Over a Network.
- DRAO: risk averse as easy as risk neutral
- DRAO-S: can be efficiently implemented
- They are both tight.
- Paper link: **Optimal Methods for Risk Averse Distributed Optimization**
- https://arxiv.org/abs/2203.05117