Distributed online optimization over jointly connected digraphs

David Mateos-Núñez   Jorge Cortés

University of California, San Diego

{*dmateosn,cortes*}*@ucsd.edu*

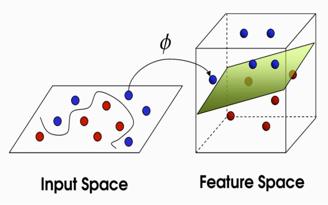
Southern California Optimization Day
UC San Diego, May 23, 2014

Overview: Distributed online optimization

Distributed optimization

Online optimization

Case study: medical diagnosis



Distributed online optimization

Why distributed?

- **information** *is distributed across group of agents*
- *need to* **interact** *to optimize performance*

Why online?

- **information** *becomes incrementally available*
- *need* **adaptive** *solution*

Machine learning in healthcare

Medical findings, symptoms:
age factor
amnesia before impact
deterioration in GCS score
open skull fracture
loss of consciousness
vomiting

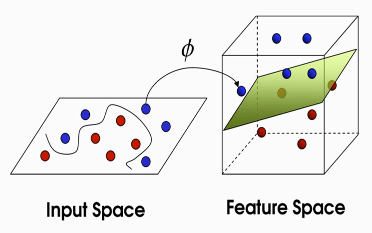
Any acute brain finding revealed on Computerized Tomography?

(-1 = not present, 1 = present)

"The Canadian CT Head Rule for patients with minor head injury"

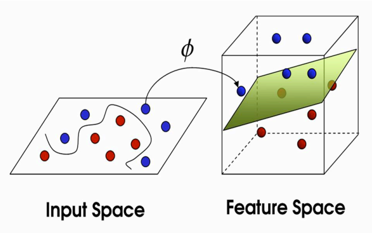
Binary classification

- feature vector of patient s: $w_s = ((w_s)_1, \ldots, (w_s)_{d-1})$
- true diagnosis: $y_s \in \{-1, 1\}$
- wanted weights: $x = (x_1, \ldots, x_d)$
- predictor: $h(x, w_s) = x^\top (w_s, 1)$
- margin: $m_s(x) = y_s \, h(x, w_s)$



Binary classification

- feature vector of patient s: $w_s = ((w_s)_1, \ldots, (w_s)_{d-1})$
- true diagnosis: $y_s \in \{-1, 1\}$
- wanted weights: $x = (x_1, \ldots, x_d)$
- predictor: $h(x, w_s) = x^\top (w_s, 1)$
- margin: $m_s(x) = y_s\, h(x, w_s)$



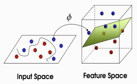
Given the data set $\{w_s\}_{s=1}^P$, estimate $x \in \mathbb{R}^d$ by solving

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \sum_{s=1}^{P} l(m_s(x))$$

where the loss function $l : \mathbb{R} \to \mathbb{R}$ is decreasing and **convex**

Review of

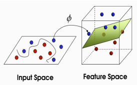
distributed convex optimization



In the diagnosis example

Health center $i \in \{1, \ldots, N\}$ manages a set of patients $\mathcal{P}^i$

$$f(x) = \sum_{i=1}^{N} \sum_{s \in \mathcal{P}^i} l(y_s h(x, w_s)) = \sum_{i=1}^{N} f^i(x)$$



In the diagnosis example

Health center $i \in \{1, \ldots, N\}$ manages a set of patients $\mathcal{P}^i$

$$f(x) = \sum_{i=1}^{N} \sum_{s \in \mathcal{P}^i} l(y_s h(x, w_s)) = \sum_{i=1}^{N} f^i(x)$$

Goal: best predicting model $w \mapsto h(x, w)$

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{N} f^i(x)$$

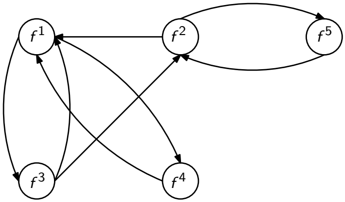
using **“local information”**

What do we mean by **"using local information"**?

Agent i maintains an estimate x_t^i of

$$x^* \in \arg\min_{x \in \mathbb{R}^d} \sum_{i=1}^{N} f^i(x)$$

- Agent i has access to f^i
- Agent i can share its estimate x_t^i with "neighboring" agents



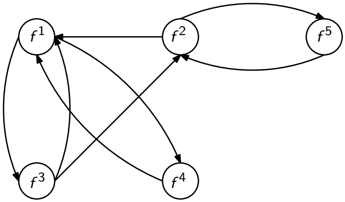
$$A = \begin{bmatrix} & & a_{13} & a_{14} & \\ a_{21} & & & & a_{25} \\ a_{31} & a_{32} & & & \\ a_{41} & & & & \\ & a_{52} & & & \end{bmatrix}$$

What do we mean by **"using local information"**?

Agent i maintains an estimate x_t^i of

$$x^* \in \arg\min_{x \in \mathbb{R}^d} \sum_{i=1}^{N} f^i(x)$$

- Agent i has access to f^i
- Agent i can share its estimate x_t^i with "neighboring" agents



$$A = \begin{bmatrix} & & a_{13} & a_{14} & \\ a_{21} & & & & a_{25} \\ a_{31} & a_{32} & & & \\ a_{41} & & & & \\ & a_{52} & & & \end{bmatrix}$$

Application to **distributed estimation in wireless sensor networks** and beyond... sensor is any channel for the machine to "learn"

How do agents agree on the optimizer?

- ⋆ Spreading of information
  (gossip, time-varying topologies, B-joint connectivity)
- ⋆ Relation between **consensus** & **local minimization**

How do agents agree on the optimizer?

⋆ Spreading of information
(gossip, time-varying topologies, B-joint connectivity)

⋆ Relation between **consensus** & **local minimization**

- **A. Nedić and A. Ozdaglar, TAC, 09**

$$z^i_{k+1} = \sum_{j=1}^{N} \mathsf{a}_{ij,k}\, z^j_k - \eta_t g, \quad x^i_{k+1} = \Pi_{\mathcal{X}}(z^i_{k+1}),$$

where $\mathsf{A}_k = (\mathsf{a}_{ij,k})$ is doubly stochastic and $g \in \partial f^i(x^i_k)$

- **J. C. Duchi, A. Agarwal, and M. J. Wainwright, TAC, 12**

Saddle-point dynamics

The minimization problem can be regarded as

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{N} f^i(x) \; = \min_{\substack{x^1,\dots,x^N \in \mathbb{R}^d \\ x^1 = \dots = x^N}} \sum_{i=1}^{N} f^i(x^i) \; = \min_{\substack{\mathbf{x} \in (\mathbb{R}^d)^N \\ \mathbf{L}\mathbf{x} = 0}} \sum_{i=1}^{N} f^i(x^i),$$

where $(\mathbf{L}\mathbf{x})^i = \sum_{j=1}^{N} \mathsf{a}_{ij}(x^i - x^j)$

Saddle-point dynamics

The minimization problem can be regarded as

$$\min_{x\in\mathbb{R}^d}\sum_{i=1}^N f^i(x) \;=\; \min_{\substack{x^1,\dots,x^N\in\mathbb{R}^d\\ x^1=\dots=x^N}}\;\sum_{i=1}^N f^i(x^i) \;=\; \min_{\substack{\mathbf{x}\in(\mathbb{R}^d)^N\\ \mathbf{L}\mathbf{x}=0}}\sum_{i=1}^N f^i(x^i),$$

where $(\mathbf{L}\mathbf{x})^i = \sum_{j=1}^N \mathsf{a}_{ij}(x^i - x^j)$

The augmented **Lagrangian** when $\mathbf{L}$ is symmetric is

$$F(\mathbf{x},\mathbf{z}) := \tilde{f}(\mathbf{x}) + \tfrac{a}{2}\,\mathbf{x}^\top\mathbf{L}\,\mathbf{x} + \mathbf{z}^\top\mathbf{L}\,\mathbf{x},$$

which is **convex-concave**, and the **saddle-point dynamics**

$$\begin{aligned}\dot{\mathbf{x}} &= -\frac{\partial F(\mathbf{x},\mathbf{z})}{\partial \mathbf{x}} = -\nabla\tilde{f}(\mathbf{x}) - a\,\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{z}\\ \dot{\mathbf{z}} &= \frac{\partial F(\mathbf{x},\mathbf{z})}{\partial \mathbf{z}} = \mathbf{L}\mathbf{x}\end{aligned}$$

Weight-balanced digraphs

$$\dot{\mathbf{x}} = \nabla\tilde{f}(\mathbf{x}) - a\,\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{z}$$
$$\dot{\mathbf{z}} = \mathbf{L}\mathbf{x}$$

Lagrangian $F(\mathbf{x},\mathbf{z}) \triangleq \tilde{f}(\mathbf{x}) + \frac{a}{2}\,\mathbf{x}^\top\mathbf{L}\,\mathbf{x} + \mathbf{z}^\top\mathbf{L}\,\mathbf{x},\qquad \tilde{f}(\mathbf{x}) = \sum_{i=1}^{N} f^i(x^i)$

Weight-balanced digraphs

$$\dot{\mathbf{x}} = \nabla \tilde{f}(\mathbf{x}) - a\,\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{z}$$
$$\dot{\mathbf{z}} = \mathbf{L}\mathbf{x}$$

Lagrangian $F(\mathbf{x},\mathbf{z}) \triangleq \tilde{f}(\mathbf{x}) + \frac{a}{2}\,\mathbf{x}^\top \mathbf{L}\,\mathbf{x} + \mathbf{z}^\top \mathbf{L}\,\mathbf{x}$, $\quad \tilde{f}(\mathbf{x}) = \sum_{i=1}^N f^i(x^i)$

$$\dot{\mathbf{x}} = -\frac{\partial F(\mathbf{x},\mathbf{z})}{\partial \mathbf{x}} = -\nabla \tilde{f}(\mathbf{x}) - a\,\tfrac{1}{2}(\mathbf{L} + \mathbf{L}^\top)\mathbf{x} - \mathbf{L}^\top \mathbf{z}$$ **(Non distributed!)**

$$\dot{\mathbf{z}} = \frac{\partial F(\mathbf{x},\mathbf{z})}{\partial \mathbf{z}} = \mathbf{L}\mathbf{x}$$

J. Wang and N. Elia (with $\mathbf{L}^\top = \mathbf{L}$), Allerton, 10

Weight-balanced digraphs

$$\dot{\mathbf{x}} = \nabla \tilde{f}(\mathbf{x}) - a\,\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{z}$$
$$\dot{\mathbf{z}} = \mathbf{L}\mathbf{x}$$

Lagrangian $F(\mathbf{x},\mathbf{z}) \triangleq \tilde{f}(\mathbf{x}) + \frac{a}{2}\,\mathbf{x}^\top \mathbf{L}\,\mathbf{x} + \mathbf{z}^\top \mathbf{L}\,\mathbf{x}, \qquad \tilde{f}(\mathbf{x}) = \sum_{i=1}^{N} f^i(x^i)$

$$\dot{\mathbf{x}} = -\frac{\partial F(\mathbf{x},\mathbf{z})}{\partial \mathbf{x}} = -\nabla \tilde{f}(\mathbf{x}) - a\,\tfrac{1}{2}\big(\mathbf{L} + \mathbf{L}^\top\big)\mathbf{x} - \mathbf{L}^\top \mathbf{z}$$ **(Non distributed!)**

changed to $-\nabla \tilde{f}(\mathbf{x}) - a\,\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{z}$

$$\dot{\mathbf{z}} = \frac{\partial F(\mathbf{x},\mathbf{z})}{\partial \mathbf{z}} = \mathbf{L}\mathbf{x}$$

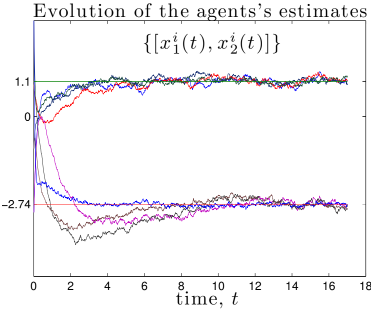
B. Gharesifard and J. Cortés, CDC, 12

Note: $a > 0$; otherwise the linear part of the saddle-point dynamics is a Hamiltonian system

Weight-balanced digraphs

$$\dot{\mathbf{x}} = \nabla \tilde{f}(\mathbf{x}) - a\,\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{z}$$
$$\dot{\mathbf{z}} = \mathbf{L}\mathbf{x}$$

Example: 4 agents in a directed cycle

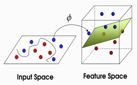


$$f_1(x_1, x_2) = \tfrac{1}{2}((x_1 - 4)^2 + (x_2 - 3)^2)$$
$$f_2(x_1, x_2) = x_1 + 3x_2 - 2$$
$$f_3(x_1, x_2) = \log(e^{x_1+3} + e^{x_2+1})$$
$$f_4(x_1, x_2) = (x_1 + 2x_2 + 5)^2 + (x_1 - x_2 - 4)^2$$

- convergence to a neighborhood of optimizer $(1.10, -2.74)$
- size of the neighborhood depends on **size of the noise** [DMN-JC, 13]

Review of
online convex optimization

Different kind of optimization: sequential decision making

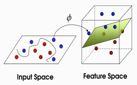


Resuming the diagnosis example:

Each round $t \in \{1, \ldots, T\}$	
question (features, medical findings):	w_t
decision (about using CT):	$h(x_t, w_t)$
outcome (by CT findings/follow up of patient):	y_t
loss:	$l(y_t\, h(x_t, w_t))$

Choose x_t **&** Incur **loss** $f_t(x_t) := l(y_t\, h(x_t, w_t))$

Different kind of optimization: sequential decision making



Resuming the diagnosis example:

Each round $t \in \{1, \dots, T\}$	
question (features, medical findings):	w_t
decision (about using CT):	$h(x_t, w_t)$
outcome (by CT findings/follow up of patient):	y_t
loss:	$l(y_t\, h(x_t, w_t))$

Choose x_t **&** Incur **loss** $f_t(x_t) := l(y_t\, h(x_t, w_t))$

Goal: sublinear regret

$$\mathcal{R}(u, T) := \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(u) \leq o(T)$$

using **"historical observations"**

Why regret?

If the regret is sublinear,

$$\sum_{t=1}^{T} f_t(x_t) \leq \sum_{t=1}^{T} f_t(u) + o(T),$$

then,

$$\frac{1}{T}\sum_{t=1}^{T} f_t(x_t) \leq \frac{1}{T}\sum_{t=1}^{T} f_t(u) + \frac{o(T)}{T}$$

In temporal average, **online** decisions $\{x_t\}_{t=1}^{T}$ perform **as well** as best fixed decision in **hindsight**

"No regrets, my friend"

What about generalization error?

- **Sublinear regret** does **not** imply x_{t+1} will do well with f_{t+1}
- No assumptions about sequence $\{f_t\}$; it can
  - follow an **unknown** stochastic or deterministic model,
  - or be chosen **adversarially**
- In our example, $f_t := l(y_t\, h(x_t, w_t))$.
  - If some model $w \mapsto h(x^*, w)$ explains reasonably the data in **hindsight**,
  - then the **online models** $w \mapsto h(x_t, w)$ perform just **as well in average**

What about generalization error?

- **Sublinear regret** does **not** imply x_{t+1} will do well with f_{t+1}
- No assumptions about sequence $\{f_t\}$; it can
  - follow an **unknown** stochastic or deterministic model,
  - or be chosen **adversarially**
- In our example, $f_t := l(y_t\, h(x_t, w_t))$.
  - If some model $w \mapsto h(x^*, w)$ explains reasonably the data in **hindsight**,
  - then the **online models** $w \mapsto h(x_t, w)$ perform just **as well in average**

Other applications:

- portfolio selection
- online advertisement placement
- interactive learning

Some classical results

Projected gradient descent:

$$x_{t+1} = \Pi_{\mathcal{S}}(x_t - \eta_t \nabla f_t(x_t)), \tag{1}$$

where $\Pi_{\mathcal{S}}$ is a projection onto a compact set $\mathcal{S} \subseteq \mathbb{R}^d$, & $\|\nabla f\|_2 \leq H$

Follow-the-Regularized-Leader:

$$x_{t+1} = \arg\min_{y \in \mathcal{S}} \left(\sum_{s=1}^{t} f_s(y) + \psi(y) \right)$$

Some classical results

Projected gradient descent:

$$x_{t+1} = \Pi_{\mathcal{S}}(x_t - \eta_t \nabla f_t(x_t)), \tag{1}$$

where $\Pi_{\mathcal{S}}$ is a projection onto a compact set $\mathcal{S} \subseteq \mathbb{R}^d$, & $\|\nabla f\|_2 \leq H$

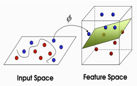
Follow-the-Regularized-Leader:

$$x_{t+1} = \arg\min_{y \in \mathcal{S}} \left(\sum_{s=1}^{t} f_s(y) + \psi(y) \right)$$

- **Martin Zinkevich**, 03
  - (1) achieves $\mathcal{O}(\sqrt{T})$ regret under convexity with $\eta_t = \frac{1}{\sqrt{t}}$
- **Elad Hazan, Amit Agarwal, and Satyen Kale**, 07
  - (1) achieves $\mathcal{O}(\log T)$ regret under p-strong convexity with $\eta_t = \frac{1}{pt}$
  - Others: Online Newton Step, Follow the Regularized Leader, etc.

Our contribution:

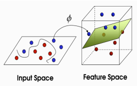
Combining both aspects



Back again to the diagnosis example:

Health center $i \in \{1, \ldots, N\}$ takes care of a set of patients $\mathcal{P}_t^i$ at time t

$$f^i(x) = \sum_{t=1}^{T} \sum_{s \in \mathcal{P}_t^i} l(y_s h(x, w_s)) = \sum_{t=1}^{T} f_t^i(x)$$



Back again to the diagnosis example:

Health center $i \in \{1, \ldots, N\}$ takes care of a set of patients $\mathcal{P}_t^i$ at time t

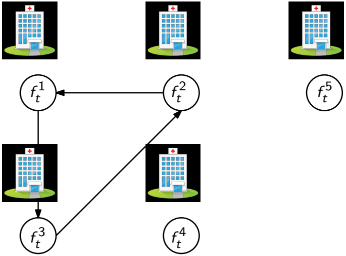
$$f^i(x) = \sum_{t=1}^{T} \sum_{s \in \mathcal{P}_t^i} l(y_s h(x, w_s)) = \sum_{t=1}^{T} f_t^i(x)$$

Goal: sublinear agent regret

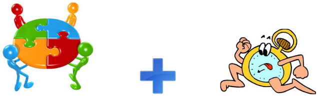
$$\mathcal{R}^j(u, T) := \sum_{t=1}^{T} \sum_{i=1}^{N} f_t^i(x_t^j) - \sum_{t=1}^{T} \sum_{i=1}^{N} f_t^i(u) \leq o(T)$$

using **“local information”** **&** **“historical observations”**

Challenge: Coordinate hospitals



Need to design distributed online algorithms



Previous work on consensus-based online algorithms

- **F. Yan, S. Sundaram, S. V. N. Vishwanathan and Y. Qi**, TAC
  Projected Subgradient Descent
  - $\log(T)$ regret (**local strong convexity** & bounded subgradients)
  - $\sqrt{T}$ regret (**convexity** & bounded subgradients)
  - Both analysis require a projection onto a **compact set**
- **S. Hosseini, A. Chapman and M. Mesbahi**, CDC, 13
  Dual Averaging
  - $\sqrt{T}$ regret (**convexity** & bounded subgradients)
  - General regularized projection onto a convex closed set.
- **K. I. Tsianos and M. G. Rabbat**, arXiv, 12
  Projected Subgradient Descent
  - Empirical risk as opposed to regret analysis

Communication digraph in all cases is **fixed, strongly connected** & weight-balanced

Our contributions (informally)

- **time-varying** communication digraphs under B-**joint connectivity** & weight-balanced
- unconstrained optimization (**no projection** step onto a **bounded** set)
- $\log T$ regret (**local strong convexity** & bounded subgradients)
- $\sqrt{T}$ regret (**convexity** & bounded subgradients)

Coordination algorithm

$$x_{t+1}^i = x_t^i \; - \eta_t \, g_{x_t^i}$$

- **Subgradient descent** on **previous local** objectives, $\quad g_{x_t^i} \in \partial f_t^i$

Coordination algorithm

$$x_{t+1}^i = x_t^i - \eta_t \, g_{x_t^i} + \sigma\Big(a \sum_{j=1, j\neq i}^{N} \mathsf{a}_{ij,t}\,(x_t^j - x_t^i)\Big)$$

- **Proportional** (linear) feedback on **disagreement** with neighbors

Coordination algorithm

$$x_{t+1}^i = x_t^i \; - \eta_t \, g_{x_t^i} \\ \qquad + \sigma\Big(a \sum_{j=1, j\neq i}^{N} \mathsf{a}_{ij,t}\big(x_t^j - x_t^i\big) + \sum_{j=1, j\neq i}^{N} \mathsf{a}_{ij,t}\big(z_t^j - z_t^i\big)\Big)$$

$$z_{t+1}^i = z_t^i \; - \sigma \sum_{j=1, j\neq i}^{N} \mathsf{a}_{ij,t}\big(x_t^j - x_t^i\big)$$

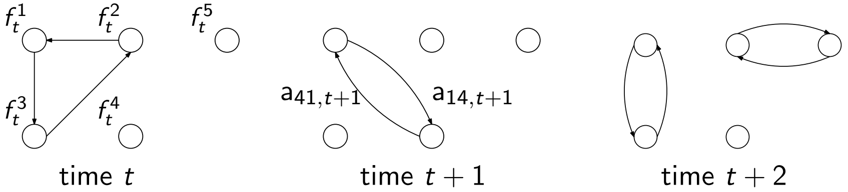
- **Integral** (linear) feedback on **disagreement** with neighbors

Coordination algorithm

$$
\begin{aligned}
x^i_{t+1} = x^i_t \; &- \eta_t \, g_{x^i_t} \\
&\qquad + \sigma\Big(a \sum_{j=1, j\neq i}^{N} \mathsf{a}_{ij,t}\big(x^j_t - x^i_t\big) + \sum_{j=1, j\neq i}^{N} \mathsf{a}_{ij,t}\big(z^j_t - z^i_t\big)\Big)
\end{aligned}
$$

$$
z^i_{t+1} = z^i_t \; - \sigma \sum_{j=1, j\neq i}^{N} \mathsf{a}_{ij,t}\big(x^j_t - x^i_t\big)
$$

- **Union** of graphs over intervals of length B is **strongly connected**.



Coordination algorithm

$$x_{t+1}^i = x_t^i \; - \eta_t \, g_{x_t^i}$$
$$+ \sigma\Big(a \sum_{j=1, j\neq i}^{N} \mathrm{a}_{ij,t}\big(x_t^j - x_t^i\big) + \sum_{j=1, j\neq i}^{N} \mathrm{a}_{ij,t}\big(z_t^j - z_t^i\big)\Big)$$
$$z_{t+1}^i = z_t^i \; - \sigma \sum_{j=1, j\neq i}^{N} \mathrm{a}_{ij,t}\big(x_t^j - x_t^i\big)$$

- Compact representation

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{z}_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{bmatrix} - \sigma \begin{bmatrix} a\mathbf{L}_t & \mathbf{L}_t \\ -\mathbf{L}_t & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{bmatrix} - \eta_t \begin{bmatrix} \tilde{g}_{\mathbf{x}_t} \\ 0 \end{bmatrix}$$

Coordination algorithm

$$x_{t+1}^i = x_t^i - \eta_t \, g_{x_t^i} + \sigma\Big(a \sum_{j=1, j\neq i}^{N} a_{ij,t}(x_t^j - x_t^i) + \sum_{j=1, j\neq i}^{N} a_{ij,t}(z_t^j - z_t^i)\Big)$$

$$z_{t+1}^i = z_t^i - \sigma \sum_{j=1, j\neq i}^{N} a_{ij,t}(x_t^j - x_t^i)$$

- Compact representation & generalization

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{z}_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{bmatrix} - \sigma \begin{bmatrix} a\mathbf{L}_t & \mathbf{L}_t \\ -\mathbf{L}_t & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{bmatrix} - \eta_t \begin{bmatrix} \tilde{g}_{\mathbf{x}_t} \\ 0 \end{bmatrix}$$

$$\mathbf{v}_{t+1} = (\mathrm{I} - \sigma G \otimes \mathbf{L}_t)\mathbf{v}_t - \eta_t \mathbf{g}_t,$$

Our contributions

Theorem

Assume that

- $\{f_t^1, ..., f_t^N\}_{t=1}^T$ *are convex functions in* $\mathbb{R}^d$
  - *with* H*-bounded subgradient sets,*
  - *nonempty and uniformly bounded sets of minimizers, and*
  - p*-strongly convex in a suff. large neighborhood of their minimizers*
- *The sequence of weight-balanced communication digraphs is*
  - *nondegenerate, and*
  - B*-jointly-connected*
- $G \in \mathbb{R}^{K \times K}$ *is diagonalizable with positive real eigenvalues*

Then, taking learning rates $\eta_t = \frac{1}{p\,t}$,

$$\mathcal{R}^j(u, T) \leq C(\|u\|_2^2 + 1 + \log T),$$

for any $j \in \{1, \ldots, N\}$ and $u \in \mathbb{R}^d$

Our contributions

Theorem

Assume that

- $\{f_t^1, ..., f_t^N\}_{t=1}^T$ *are convex functions in* $\mathbb{R}^d$
  - *with H-bounded subgradient sets,*
  - *nonempty and uniformly bounded sets of minimizers, and*
  - *p-strongly convex in a suff. large neighborhood of their minimizers*
- *The sequence of weight-balanced communication digraphs is*
  - *nondegenerate, and*
  - *B-jointly-connected*
- $G \in \mathbb{R}^{K \times K}$ *is diagonalizable with positive real eigenvalues*

Relaxing strong convexity to **convexity** and using the Doubling Trick scheme (see S. Shalev-Shwartz) for the learning rates,

$$\mathcal{R}^j(u, T) \leq C\|u\|_2^2\sqrt{T},$$

for any $j \in \{1, \ldots, N\}$ and $u \in \mathbb{R}^d$

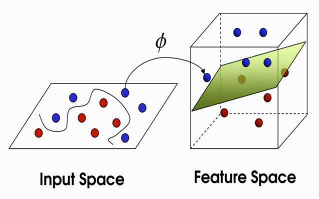
Idea of the proof

- Network regret

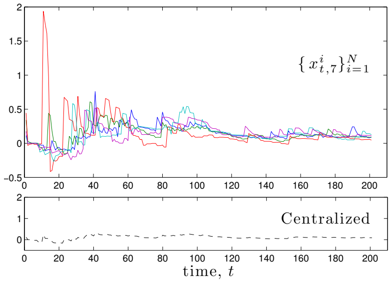
$$\mathcal{R}^j(u, T) := \sum_{t=1}^{T} \sum_{i=1}^{N} f_t^i(x_t^i) - \sum_{t=1}^{T} \sum_{i=1}^{N} f_t^i(u)$$

- Disagreement dynamics under B-joint connectivity
- Bound on the trajectories uniform in T

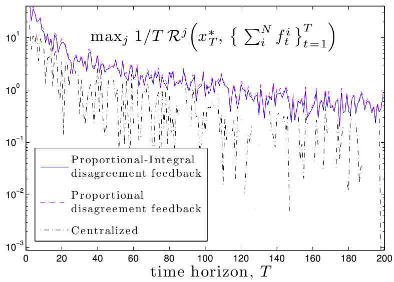
Simulations:
acute brain finding revealed on Computerized Tomography



Agents' estimates



Average regret



$$f_t^i(x) = \sum_{s \in \mathcal{P}_t^i} l(y_s h(x, w_s)) + \frac{1}{10}\|x\|_2^2$$

where

$$l(m) = \log\left(1 + e^{-2m}\right)$$

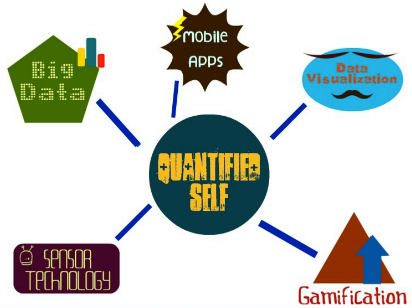
Conclusions

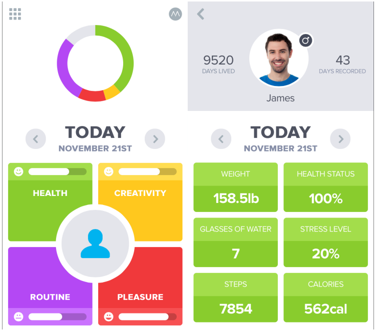
- Distributed online unconstrained convex optimization with **sublinear regret** under B-joint connectivity
- Relevant for **regression & classification** that play a crucial role in machine learning, computer vision, etc.

Future work

- Refine guarantees under **model for evolution** of objective functions
- Enable agents to cooperatively **select features** that strike the balance sensibility/specificity
- Effect of **noise** on the performance

Future horizons for distributed optimization in healthcare





Engage & detect disease before it happens

Thank you for listening!