Distributed / Multi-agent optimization

- we have a collection of N agents
- optimization problem: $\min_{x \in \mathbb{R}^n} \sum_{i=1}^{N} f_i(x)$,

where f_i is the cost function of i-th agent

- f_i is only known to agent i
- agents are unwilling to share cost function f_i with others.

- Example: Suppose we wish to solve: $\min_{w} \sum_{i=1}^{D} (w^T x_i - y_i)^2$,

$(x_i, y_i)_{i=1}^{D}$ are collectively held by N agents.

agent 1 holds $(x_i, y_i)_{i=1}^{D_1}$

agent 2 holds $(x_i, y_i)_{i=D_1+1}^{D_1+D_2}$

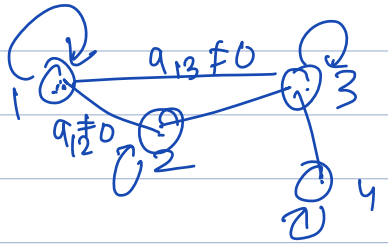
$\vdots$

agent N holds $(x_i, y_i)_{i=D_1+D_2+\cdots+D_{N-1}+1}^{D}$

$$\min_{w} \left[\underbrace{\sum_{i=1}^{D_1} (w^T x_i - y_i)^2}_{f_1(w)} + \underbrace{\sum_{i=D_1+1}^{D_2} (\quad)^2}_{f_2(w)} + \cdots + (\quad) \right]$$

Communication graph: $G = (V, E)$

nodes, edges



$|V| = N$, each agent is a node

$(v_i, v_j) \in E$ if agent i & j can exchange information

the weight of this edge a_{ij}

$$a_{ij} = \begin{cases} 0 & \text{if } (v_i, v_j) \notin E \\ > 0 & \text{if } (v_i, v_j) \in E \end{cases}$$

we will consider graphs that are connected & undirected.

$a_{ij} = a_{ji}$

for node i: we define $\mathcal{N}_i$ to be its set of neighbors

$$\mathcal{N}_i = \{ v_j \in V \mid a_{ij} \neq 0 \}$$

Distributed Gradient Descent

- each agent initializes a solution x_0^i
- at every time $t = 0, 1, 2, \ldots$
  - gather: x_t^j from $j \in \mathcal{N}_i$
  - compute: $v_{t+1}^i = \sum_{j \in \mathcal{N}_i} a_{ij} x_t^j$
  - update: $x_{t+1}^i = v_{t+1}^i - \eta_t \nabla f_i(v_{t+1}^i)$

Assumptions

i) on the network: $a_{ii} > 0 \ \forall i$, $\sum_{i=1}^{N} a_{ij} = 1 \ \forall j$, $\sum_{j=1}^{N} a_{ij} = 1 \ \forall i$

$A\mathbb{1} = \mathbb{1}$

$\mathbb{1}^T A = \mathbb{1}^T$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots \\ a_{21} & a_{22} & \cdots \\ a_{n1} & a_{n2} & \cdots a_{nn} \end{bmatrix}$$

ii) step-sizes: $\eta_t \geq 0$, $\sum_{t=0}^{\infty} \eta_t = \infty$, $\sum_{t=0}^{\infty} \eta_t^2 < \infty$

iii) on the cost functions:
- each f_i is convex, $\|\nabla f_i(x)\| \leq C_i \ \forall i$
- $\min_x \sum_{i=1}^{N} f_i(x)$ has at least one optimal solution.

Theorem: The solutions generated by the distributed GD algorithm converges to an optimal solution x^* for every agent, i.e.,

$$\lim_{t \to \infty} \|x_t^i - x^*\| = 0 \quad \forall\, i \in \{1, 2 \dots N\}.$$

proof sketch:

i) consensus: let $\bar{x}_t = \frac{1}{N} \sum_{i=1}^{N} x_t^i$

$$\lim_{t \to \infty} \|x_t^i - \bar{x}_t\| = 0 \quad \forall\, i.$$

ii) $$\lim_{t \to \infty} \sum_{t=0}^{T} \eta_t \|x_t^i - \bar{x}_t\| < \infty \quad \forall\, i$$

iii) $$\lim_{t \to \infty} \|\bar{x}_t - x^*\| = 0.$$

Note: we cannot use constant step-size because each agent is only computing partial gradient, rather than true gradient. This leads to slow convergence.

Gradient Tracking Algorithm

$x_{t+1} = x_t - \eta \nabla f(x_t)$

- Initialize: x_0^i, $y_0^i = \nabla f_i(x_0^i)$
- At each time t
  - gather: x_t^j from $j \in N_i$
  - updates: $x_{t+1}^i = \sum_{j \in N_i} a_{ij} x_t^j - \eta y_t^i$
  - gather: y_t^j from $j \in N_i$

- updates :- $y_{t+1}^i = \sum_{j \in N_i} a_{ij} y_t^j + \left[\nabla f_i(x_{t+1}^{(i)}) - \nabla f_i(x_t^{(i)}) \right]$

privately held info agent i

Assumptions:

i) each f_i is β-smooth & strongly convex

ii) step-size $\eta < \bar{\eta}$, where $\bar{\eta} \in \left(0, \frac{N}{\beta}\right)$

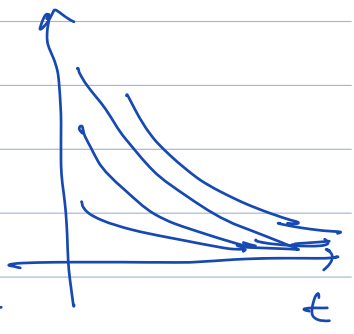
Theorem : $\|x_{t+1}^i - x^*\| \le \rho \|x_t^i - x^*\|$ for $\rho \in (0,1)$.

$\Rightarrow \|x_T^i - x^*\| \le \rho^T \|x_0^i - x^*\|$

Performance of these algorithms are evaluated as follows.

- consensus : $e_t^i = \|x_t^i - \bar{x}_t\|$ vs. t
- optimal value : $|f(\bar{x}_t) - f(x^*)|$ vs. t
- optimal solution : $\|\bar{x}_t - x^*\|$ vs. t

$\max_i \|x_t^i - x^*\|$ vs. t



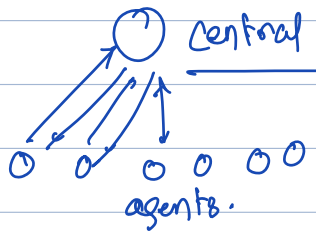
- log-scale

Many extensions :- time-varying graphs, directed graph, asynchronous communication, malicious agents, accelerated schemes, and so on.

20th March

Local Stochastic Gradient Descent (Local SGD)

- decentralized algorithm: there is a central server which helps agents to converge to the optimal solution.

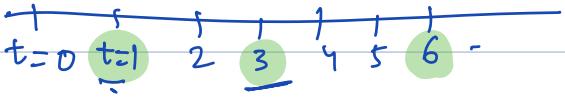


- consider the problem: $\min_{x \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^{N} f_i(x)$,

where $f_i(x) = \sum_{j=1}^{N_i} \ell_j^i(x, \xi_j^i)$,

with $(\xi_j^i)_{j=1}^{N_i}$ is the data held by agent i

- Local SGD
  - Let $I_T \subseteq \{1, 2, \dots T\}$ be the time stamps at which communication takes place.
  - every agents starts with x_0^i
  - at very time t:



$$x_{t+1}^i = \begin{cases} x_t^i - \eta_t \nabla \ell_{i_t}^i(x_t^i) & \text{if } t \notin I_T, \ i_t \in \{1, 2, \dots N_i\} \\ \frac{1}{N} \sum_{k=1}^{N} \left[x_t^k - \eta_t \nabla \ell_{k_t}^k(x_t^k) \right] & \text{if } t \in I_T \end{cases}$$

→ each agent i evaluates $[\cdot]$
→ communicates with central server
→ server takes avg., & sends it back.

Theorem Let $\eta^t = \frac{4}{\beta(a+t)}$, where a is a constant & β is the smoothness parameter of the function.

Let the function be α-strongly convex as well.

Then $$\mathbb{E}\left[f\left(\frac{1}{N}\sum_{i=1}^{N} x_T^i\right)\right] - f(x^*) \leq \mathcal{O}\left(\frac{1}{KT}\right),$$

where $K = \frac{\beta}{\alpha}$ is the condition number.

(Q) What to do when the function is not differentiable?

In those settings, we make use of subgradients.

Defn: For a function f, the set of subgradients at x_0 is defined as

$$\partial f(x_0) = \{ g \in \mathbb{R}^n \mid f(y) \geq f(x_0) + g^T(y - x_0) \ \forall y \in \mathbb{R}^n \}.$$

$\partial f(x_0)$ is called the subdifferential.

- In general $\partial f(x_0)$ may be an empty-set.
- However, if the function is convex, then $\partial f(x_0)$ is non-empty.

- If f is differentiable at x_0, then $\partial f(x_0) = \{\nabla f(x_0)\}$.

Example:

$f(x) = |x|$.

$\partial f(1) = 1$

$\partial f(0) = \{[-1, 1]\}$

For such functions, we use subgradient (desent) algorithms:

$$x_{t+1} = x_t - \eta_t g_t, \text{ where } g_t \in \partial f(x_t).$$

Example: $f(x) = \sum_{i=1}^{N} |x_i| = \|x\|_1, \quad x \in \mathbb{R}^n$

find $\partial f\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right) = \{ g \in \mathbb{R}^n \mid g_i \in [-1, 1], \forall i \}$.

we want to show that if $w \in [-1, 1]^n$, then

$$f(y) \geq f(x) + w^T (y - x) \quad \forall\, y \in \mathbb{R}^n \quad \text{at} \quad x = 0.$$

$$\Rightarrow \quad |y|_1 \geq w^T y$$

from holder's inequality: $w^T y \leq \|w\|_\infty \|y\|_1 \leq \|y\|_1$

$$\|w\|_\infty = \max_{1 \leq i \leq N} |w_i|$$

- These functions are not smooth, but may be strongly convex.
- Under the assumption that $\|g(x)\| \leq G$, then we can establish convergence of subgradient descent in an analogous manner to the convergence of gradient descent under bounded gradient assumption.

Application: Optimal Detection.

Let $\Theta = \{1, 2, \ldots m\}$ be the set of hypothesis.

$Y = \{1, 2, \ldots n\}$ be the set of observations.

eg: $\Theta = \{$common cold, covid-1, covid-2, flu$\}$

$Y = \{$fever, sneezing, sore throat, fever+sore throat, fever+sneezing$\}$

For every hypothesis, we have a distribution over Y.

Let p_i denote the conditional distribution over Y when $\Theta = i$

$$p_i = \begin{bmatrix} \mathbb{P}(Y=1 \mid \Theta=i) \\ \mathbb{P}(Y=2 \mid \Theta=i) \\ \vdots \\ \mathbb{P}(Y=n \mid \Theta=i) \end{bmatrix} \in \mathbb{R}^n, \quad p_i \geq 0, \ \mathbb{1}^T p_i = 1$$

p_i are known

$$P = [p_1 \ p_2 \ \cdots \ p_m]$$

Question: Given an observation Y, estimate (distribution) over Θ.

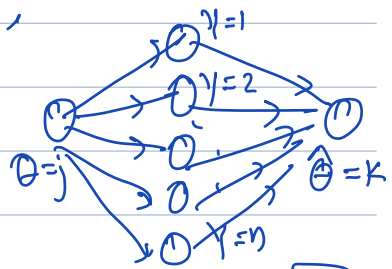
For every observation, we have a different distribution over Θ.

Let the distribution for $Y = i$ be denoted as $t_i \in \mathbb{R}^m$

$$t_i = \begin{bmatrix} \mathbb{P}(\hat{\Theta}=1 \mid Y=i) \\ \mathbb{P}(\hat{\Theta}=2 \mid Y=i) \\ \vdots \\ \mathbb{P}(\hat{\Theta}=m \mid Y=i) \end{bmatrix}, \quad t_i \geq 0, \ \mathbb{1}^T t_i = 1$$

$$T = [t_1 \ t_2 \ \cdots \ t_n]$$

we need to determine $(t_1, t_2 \cdots t_n)$,



How to evaluate the performance of t_i?

Let the ground truth be $\Theta = j$.

we want $(t_1, t_2, \cdots t_n)$ be such that $\begin{bmatrix} \mathbb{P}(\hat{\Theta}=j \mid \Theta=j) \text{ is high} \\ \& \ \mathbb{P}(\hat{\Theta} \neq j \mid \Theta=j) \text{ is low.} \end{bmatrix}$

$$\mathbb{P}(\hat{\Theta}=K \mid \Theta=j) = \sum_{i=1}^{N} \mathbb{P}(Y=i \mid \Theta=j) \cdot \mathbb{P}(\hat{\Theta}=K \mid Y=i)$$

$$= \sum_{i=1}^{N} (p_j)_i \ (t_i)_K$$

$= [TP]_{kj}$ → multiplication of k^{th} row of T with j^{th} column of P

Thus, the matrix TP has the property that

$$[TP]_{ij} = P(\hat{\Theta} = i \mid \Theta = j)$$

TP → detection matrix.

we want $[TP]_{ii}$ should be high

$[TP]_{ij}$ with $i \neq j$ should be small.

Note that:

$$[TP]_{ii} + \sum_{j \neq i} [TP]_{ij} = 1$$

$$\Rightarrow 1 - [TP]_{ii} = \sum_{j \neq i} [TP]_{ij}$$

We can now formulate the following optimization problem.

$$\begin{cases} \min\limits_{t_1, t_2 \cdots t_n} \ \max\limits_{1 \le i \le n} \left(1 - [TP]_{ii}\right) \\ \text{s.t.} \quad t_i \ge 0, \ \sum t_i = 1 \quad \forall\, i = 1, 2 \cdots n \end{cases}$$

min-max detector

Is the above problem convex??

- constraints: convex set
- cost function: it needs to be shown that $\forall\, i$, $1 - [TP]_{ii}$ is a convex function.

$$f_i(t_1, t_2 \cdots t_n) = 1 - [TP]_{ii}$$

$$= 1 - \sum_{k=1}^{n} t_{ik} p_{ki} \quad : \text{affine in } t_i$$

$$(\Pi_X(y) - y)^T (y - x) \leq 0 \quad \forall x$$