Two Algorithms for Smooth Distributed Convex Optimization

Pranav Reddy

27 November 2024

Overview

Goals

1. Discuss gradient-based methods for smooth distributed convex optimization

Goals

1. Discuss gradient-based methods for smooth distributed convex optimization

2. Compare the advantages and disadvantages of different methods

Goals

1. Discuss gradient-based methods for smooth distributed convex optimization
2. Compare the advantages and disadvantages of different methods
3. Compare the distributed and centralized settings, and understand the differences

Review I

Recall the problem

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

where $f_1, \ldots, f_N$ are L-smooth, G-Lipschitz convex functions.

Review I

Recall the problem

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

where $f_1, \dots, f_N$ are L-smooth, G-Lipschitz convex functions.
We can rewrite this as

$$\begin{aligned} \min_{x_1, \dots, x_N \in \mathbb{R}^d} \quad & \frac{1}{N} \sum_{i=1}^{N} f_i(x_i) \\ \text{s.t.} \quad & x_1 = x_2 = \dots = x_N \end{aligned}$$

Review I

Recall the problem

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

where $f_1, \ldots, f_N$ are L-smooth, G-Lipschitz convex functions. We can rewrite this as

$$\min_{x_1, \ldots, x_N \in \mathbb{R}^d} \quad \frac{1}{N} \sum_{i=1}^{N} f_i(x_i)$$
$$\text{s.t.} \quad x_1 = x_2 = \cdots = x_N$$

The error in satisfying the constraint $x_1 = \cdots = x_N$ is called the **consensus error**. We usually measure $E = \sum_{k=1}^{N} \|x_k - \bar{x}\|_2^2$.

Review II

Definition

A **graph** is a pair $G = (V, E)$ of sets, called vertices and edges, where $E \subseteq V \times V$.

- In the distributed context, vertex i represents an agent, with its local cost function $f_i \colon \mathbb{R}^d \to \mathbb{R}$ and local variable $x_i \in \mathbb{R}^d$.

Review II

Definition

A **graph** is a pair $G = (V, E)$ of sets, called vertices and edges, where $E \subseteq V \times V$.

- In the distributed context, vertex i represents an agent, with its local cost function $f_i \colon \mathbb{R}^d \to \mathbb{R}$ and local variable $x_i \in \mathbb{R}^d$.
- Edge (i, j) implies that there is a communication line between agents i and j, so agents i and j are able to communicate.

Review II

Definition

A **graph** is a pair $G = (V, E)$ of sets, called vertices and edges, where $E \subseteq V \times V$.

- In the distributed context, vertex i represents an agent, with its local cost function $f_i \colon \mathbb{R}^d \to \mathbb{R}$ and local variable $x_i \in \mathbb{R}^d$.
- Edge (i, j) implies that there is a communication line between agents i and j, so agents i and j are able to communicate.
- We assume that for any pair of agents i and j, there exists a sequence of edges $(i, i_1), (i_1, i_2), \ldots, (i_k, j)$, starting with i and ending with j.

Review II

Definition

A **graph** is a pair $G = (V, E)$ of sets, called vertices and edges, where $E \subseteq V \times V$.

- In the distributed context, vertex i represents an agent, with its local cost function $f_i\colon \mathbb{R}^d \to \mathbb{R}$ and local variable $x_i \in \mathbb{R}^d$.
- Edge (i, j) implies that there is a communication line between agents i and j, so agents i and j are able to communicate.
- We assume that for any pair of agents i and j, there exists a sequence of edges $(i, i_1), (i_1, i_2), \ldots, (i_k, j)$, starting with i and ending with j.
    - This property is known as **connectedness**.

Review II

Definition

A **graph** is a pair $G = (V, E)$ of sets, called vertices and edges, where $E \subseteq V \times V$.

- In the distributed context, vertex i represents an agent, with its local cost function $f_i : \mathbb{R}^d \to \mathbb{R}$ and local variable $x_i \in \mathbb{R}^d$.
- Edge (i, j) implies that there is a communication line between agents i and j, so agents i and j are able to communicate.
- We assume that for any pair of agents i and j, there exists a sequence of edges $(i, i_1), (i_1, i_2), \ldots, (i_k, j)$, starting with i and ending with j.
  - This property is known as **connectedness**.
- We also assume that if $(i, j) \in E$ then $(j, i) \in E$. This means all communication is bidirectional.

Review II

Definition

A **graph** is a pair $G = (V, E)$ of sets, called vertices and edges, where $E \subseteq V \times V$.

- In the distributed context, vertex i represents an agent, with its local cost function $f_i : \mathbb{R}^d \to \mathbb{R}$ and local variable $x_i \in \mathbb{R}^d$.
- Edge (i, j) implies that there is a communication line between agents i and j, so agents i and j are able to communicate.
- We assume that for any pair of agents i and j, there exists a sequence of edges $(i, i_1), (i_1, i_2), \ldots, (i_k, j)$, starting with i and ending with j.
  - This property is known as **connectedness**.
- We also assume that if $(i, j) \in E$ then $(j, i) \in E$. This means all communication is bidirectional.
  - A graph with this property is called **undirected**.

Distributed Gradient Descent

Definition

Let $W \in \mathbb{R}^{N \times N}$ be a weight matrix (compatible with the graph G) with $\sigma := \left\| W - \frac{1}{N} 11^\top \right\|_2 < 1$. The **decentralized gradient descent method** is

$$x_i(t+1) = \sum_{j=1}^{N} W_{ij} x_j(t) - \eta_t \nabla f_i(x_i(t)), \qquad i = 1, \ldots, N \quad (2.1)$$

where $\eta_t > 0$ is the step size. A common variant of the DGD method is the **diffusion method**.

$$x_i(t+1) = \sum_{j=1}^{N} W_{ij} \left(x_j(t) - \eta_t \nabla f_j(x_j(t)) \right), \qquad i = 1, \ldots, N.$$

Notation

We introduce the notation

$$X(t) = \begin{bmatrix} x_1(t)^\top \\ \vdots \\ x_N(t)^\top \end{bmatrix} \in \mathbb{R}^{N \times d}.$$

We also define the function $F : \mathbb{R}^{N \times d} \to \mathbb{R}$ defined by

$$F(X) = \sum_{i=1}^{N} f_i(x_i), \quad \text{where } X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix}.$$

Notation II

Note that, since each f_i is assumed to be differentiable, the function F is also differentiable, and its gradient is given by

$$\nabla F(X) = \begin{bmatrix} \frac{\partial F(X)}{\partial X_{1,1}} & \cdots & \frac{\partial F(X)}{\partial X_{1,d}} \\ \vdots & \ddots & \vdots \\ \frac{\partial F(X)}{\partial X_{N,1}} & \cdots & \frac{\partial F(X)}{\partial X_{N,d}} \end{bmatrix} = \begin{bmatrix} \nabla f_1(x_1)^\top \\ \vdots \\ \nabla f_N(x_N)^\top \end{bmatrix},$$

where $X = \begin{bmatrix} -x_1^\top \\ \vdots \\ -x_N^\top \end{bmatrix}$.

Notation III

Therefore, the DGD method and the diffusion method can be equivalently written as

$$X(t+1) = WX(t) - \eta_t \nabla F(X(t))$$

and

$$X(t+1) = W\Big(X(t) - \eta_t \nabla F(X(t))\Big),$$

respectively. We will mainly focus on the original DGD method, but the convergence analysis and results can be adapted to the diffusion method without much difficulty.

Notation IV

We introduce one last piece of notation:

$$\bar{x}(t) = \frac{1}{N}\sum_{k=1}^{N} x_k(t) = \frac{1}{N}(\mathbf{1}^\top X(t))^\top,$$

$$\bar{g}(t) = \frac{1}{N}\sum_{k=1}^{N} \nabla f_i(x_i(t)) = \frac{1}{N}\mathbf{1}^\top \nabla F(X(t)),$$

where $\mathbf{1} \in \mathbb{R}^N$ is the vector of all 1's. We then define the consensus error as

$$E(t) = \begin{bmatrix} (x_1(t) - \bar{x})^\top \\ \vdots \\ (x_N(t) - \bar{x})^\top \end{bmatrix} = X(t) - \mathbf{1}\bar{x}^\top = \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right) X(t).$$

First Observations

Notice that if W is doubly stochastic ($W\mathbf{1} = W^\top \mathbf{1} = \mathbf{1}$), then

$$\begin{aligned}
E(t+1) &= \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right) X(t+1) \\
&= \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(WX(t) - \eta_t \nabla F(X(t))\right) \\
&= \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(E - \frac{1}{N}\mathbf{1}\mathbf{1}^\top X(t)\right) \\
&\qquad - \eta_t \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\nabla F(X(t)) \\
&= \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right) E(t) - \eta_t \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\nabla F(X(t))
\end{aligned}$$

Consensus Error

By denoting

$$\Delta(t) = \left(I - \frac{1}{N}11^\top\right)\nabla F(X(t)),$$

we directly have

$$\|E(t+1)\|_F \leq \sigma\|E(t)\|_F + \eta_t\|\Delta(t)\|_F.$$

This inequality leads to the following result:

Lemma

Suppose there exists $\delta > 0$ such that $\|\Delta(t)\| \leq \delta$ for all t. Let $\eta_0 \geq \eta_1 \geq \eta_2 \geq \cdots$ be a sequence of non-increasing step sizes. Then

$$\|E(t)\|_F \leq \sigma^t\|E(0)\|_F + \delta\sum_{\tau=0}^{t-1}\sigma^{t-1-\tau}\eta_\tau,$$

Convergence of the Consensus Error

Lemma

Suppose there exists $\delta > 0$ such that $\|\Delta(t)\| \leq \delta$ for all t. Let $\eta_0 \geq \eta_1 \geq \eta_2 \geq \cdots$ be a sequence of non-increasing step sizes. Then

$$\|E(t)\|_F \leq \sigma^t \|E(0)\|_F + \delta \sum_{\tau=0}^{t-1} \sigma^{t-1-\tau} \eta_\tau,$$

It is easy to see that the first term decays to zero, since we assume $\sigma < 1$. The second term is slightly more technical to analyze. We will use a convenient result about sequences and series, a proof can be found online or in many textbooks on the subject.

Stolz-Ces̀aro Theorem

Theorem (Stolz-Ces̀aro Theorem)

Let (a_n) and (b_n) be two sequences of real numbers such that:

1. *(b_n) is strictly monotone and unbounded, i.e., either $b_n \to \infty$ or $b_n \to -\infty$ as $n \to \infty$.*
2. *The limit*
$$\lim_{n\to\infty} \frac{a_{n+1} - a_n}{b_{n+1} - b_n}$$
exists and is finite.

Then:
$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{a_{n+1} - a_n}{b_{n+1} - b_n}.$$

Stolz-Cesàro Theorem

Theorem (Stolz-Cesàro Theorem)

Let (a_n) and (b_n) be two sequences of real numbers such that:

1. *(b_n) is strictly monotone and unbounded, i.e., either $b_n \to \infty$ or $b_n \to -\infty$ as $n \to \infty$.*
2. *The limit*

$$\lim_{n\to\infty} \frac{a_{n+1} - a_n}{b_{n+1} - b_n}$$

exists and is finite.

Then:

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{a_{n+1} - a_n}{b_{n+1} - b_n}.$$

This can be viewed as a discrete version of L'Hopital's rule.

Proof of Convergence

Lemma

Fix two constants $\sigma \in (0,1)$ and $\beta > 0$. We have

$$\lim_{n\to\infty}(1-\sigma)n^{\beta}\sum_{\tau=1}^{n}\frac{\sigma^{n-\tau}}{\tau^{\beta}} = 1.$$

Thus, the series $\sum_{\tau=1}^{n}\frac{\sigma^{n-\tau}}{\tau^{\beta}}$ converges to 0 at a rate $\mathcal{O}(n^{-\beta})$ as $n \to \infty$.

Proof of Convergence II

We first write

$$(1-\sigma)n^{\beta}\sum_{\tau=1}^{n}\frac{\sigma^{n-\tau}}{\tau^{\beta}} = \left(\frac{1}{(1-\sigma)n^{\beta}\sigma^{n}}\right)^{-1}\sum_{\tau=1}^{n}\frac{\sigma^{-\tau}}{\tau^{\beta}}.$$

and then define two new sequences

$$a_n = \sum_{\tau=1}^{n}\frac{\sigma^{-\tau}}{\tau^{\beta}}, \qquad b_n = \frac{1}{(1-\sigma)n^{\beta}\sigma^{n}}.$$

Since $0 < \sigma < 1$, we know b_n is monotonically increasing and $\lim_{n\to\infty} b_n = \infty$.

Proof of Convergence III

In addition, it is easy to verify that

$$a_{n+1} - a_n = \frac{\sigma^{-n-1}}{(n+1)^\beta}$$

and

$$\begin{aligned} b_{n+1} - b_n &= \frac{1}{(1-\sigma)(n+1)^\beta \sigma^{n+1}} - \frac{1}{(1-\sigma)n^\beta \sigma^n} \\ &= \frac{n^\beta - \sigma(n+1)^\beta}{\sigma^{n+1}(1-\sigma)n^\beta (n+1)^\beta}. \end{aligned}$$

Thus, we can verify that

$$\frac{a_{n+1} - a_n}{b_{n+1} - b_n} = \frac{\frac{\sigma^{-n-1}}{(n+1)^\beta}}{\frac{n^\beta - \sigma(n+1)^\beta}{\sigma^{n+1}(1-\sigma)n^\beta(n+1)^\beta}} = \frac{(1-\sigma)n^\beta}{n^\beta - \sigma(n+1)^\beta} \to 1.$$

Proof of Convergence IV

Thus,

$$\frac{a_{n+1} - a_n}{b_{n+1} - b_n} = \frac{\frac{\sigma^{-n-1}}{(n+1)^\beta}}{\frac{n^\beta - \sigma(n+1)^\beta}{\sigma^{n+1}(1-\sigma)n^\beta(n+1)^\beta}} = \frac{(1-\sigma)n^\beta}{n^\beta - \sigma(n+1)^\beta} \to 1.$$

By the Stolz-Cesàro Theorem,

$$\lim_{n\to\infty} = (1-\sigma)n^\beta \sum_{\tau=1}^{n} \frac{\sigma^{n-\tau}}{\tau^\beta} = \lim_{n\to\infty} a_n/b_n = 1$$

This also proves that $\sum_{\tau=1}^{n} \frac{\sigma^{n-\tau}}{\tau^\beta} = \Theta(n^{-\beta})$.

Another Useful Bound

Lemma

Suppose there exists $\delta > 0$ *such that* $\|\Delta(t)\| \leq \delta$ *for all* t*. Let* $\eta_0 \geq \eta_1 \geq \eta_2 \geq \cdots$ *be a sequence of non-increasing step sizes. Then*

$$\sum_{\tau=0}^{t-1} \eta_\tau \|E(\tau)\|_F^2 \leq \frac{2\eta_0 \|E(0)\|_F^2}{1-\sigma^2} + \frac{4\delta^2}{(1-\sigma^2)^2} \sum_{\tau=0}^{t-2} \eta_\tau^3.$$

Proof

By induction, we can derive from the earlier recursion that

$$\begin{aligned}\|E(t+1)\|_F^2 &\leq \left(1 + \frac{1-\sigma^2}{2\sigma^2}\right)\sigma^2\|E(t)\|_F^2 + \left(1 + \frac{2\sigma^2}{1-\sigma^2}\right)\eta_t^2\delta^2 \\ &\leq \frac{1+\sigma^2}{2}\|E(t)\|_F^2 + \frac{2\delta^2}{1-\sigma^2}\eta_t^2,\end{aligned}$$

where we used the inequality

$$\|u+v\|^2 \leq (1+\varepsilon)\|u\|^2 + (1+\varepsilon^{-1})\|v\|^2$$

for any $\varepsilon > 0$.

Proof II

Consequently,

$$\sum_{\tau=0}^{t-1} \eta_\tau \|E(\tau)\|_F^2 \leq \sum_{\tau=0}^{t-1} \eta_\tau \left(\frac{1+\sigma^2}{2}\right)^{\tau} \|E(0)\|_F^2 + \frac{2\delta^2}{1-\sigma^2} \sum_{\tau=0}^{t-1} \sum_{s=0}^{\tau-1} \left(\frac{1+\sigma^2}{2}\right)^{\tau-1-s} \eta_\tau \eta_s^2.$$

Now for the first term on the right-hand side, we have

$$\sum_{\tau=0}^{t-1} \eta_\tau \left(\frac{1+\sigma^2}{2}\right)^{\tau} \|E(0)\|_F^2 \leq \eta_0 \|E(0)\|_F^2 \sum_{\tau=0}^{t-1} \left(\frac{1+\sigma^2}{2}\right)^{\tau} \leq \frac{2\eta_0 \|E(0)\|_F^2}{1-\sigma^2},$$

Proof III

To bound the second term, we can interchange the double sum to get

$$\sum_{\tau=0}^{t-1}\sum_{s=0}^{\tau-1}\left(\frac{1+\sigma^2}{2}\right)^{\tau-1-s}\eta_\tau\eta_s^2 \le \sum_{s=0}^{t-2}\eta_s^2\sum_{\tau=s+1}^{t-1}\left(\frac{1+\sigma^2}{2}\right)^{\tau-1-s}\eta_\tau$$
$$\le \frac{2}{1-\sigma^2}\sum_{s=0}^{t-2}\eta_s^3.$$

Therefore,

$$\sum_{\tau=0}^{t-1}\eta_\tau\|E(\tau)\|_F^2 \le \frac{2\eta_0\|E(0)\|_F^2}{1-\sigma^2}+\frac{4\delta^2}{(1-\sigma^2)^2}\sum_{s=0}^{t-2}\eta_s^3.$$

First Observations

Notice that

$$\bar{x}(t+1) = \bar{x}(t) - \eta_t \bar{g}(t) \tag{2.2}$$

Since $\bar{g}(t) = \frac{1}{N}\sum_{k=1}^{N} \nabla f_i(x_i(t))$, we expect that if $E(t)$ is small, then the "residual"

$$\bar{g}(t) - \frac{1}{N}\sum_{k=1}^{N} \nabla f_i(\bar{x}(t)) = \frac{1}{N}\sum_{k=1}^{N} \left(\nabla f_i(x_i(t)) - \nabla f_i(\bar{x}(t)) \right)$$

should be small.

First Observations

Notice that

$$\bar{x}(t+1) = \bar{x}(t) - \eta_t \bar{g}(t) \tag{2.2}$$

Since $\bar{g}(t) = \frac{1}{N}\sum_{k=1}^{N} \nabla f_i(x_i(t))$, we expect that if $E(t)$ is small, then the "residual"

$$\bar{g}(t) - \frac{1}{N}\sum_{k=1}^{N} \nabla f_i(\bar{x}(t)) = \frac{1}{N}\sum_{k=1}^{N} \left(\nabla f_i(x_i(t)) - \nabla f_i(\bar{x}(t)) \right)$$

should be small.

- This inexactness property is common in distributed algorithms

First Observations

Notice that

$$\bar{x}(t+1) = \bar{x}(t) - \eta_t \bar{g}(t) \tag{2.2}$$

Since $\bar{g}(t) = \frac{1}{N}\sum_{k=1}^{N} \nabla f_i(x_i(t))$, we expect that if $E(t)$ is small, then the "residual"

$$\bar{g}(t) - \frac{1}{N}\sum_{k=1}^{N} \nabla f_i(\bar{x}(t)) = \frac{1}{N}\sum_{k=1}^{N} \Big(\nabla f_i(x_i(t)) - \nabla f_i(\bar{x}(t)) \Big)$$

should be small.

- This inexactness property is common in distributed algorithms
- We can only estimate the true gradient at $\bar{x}(t)$ via some averaging process.

Proof of Convergence

Note that

$$\begin{aligned}\|\bar{x}(t) - x^\star\|^2 &= \|\bar{x}(x+1) + \eta_t \bar{g}(t) - x^\star\|^2 \\ &= \|\bar{x}(t+1) - x^\star\|^2 + 2\eta_t \langle \bar{g}(t), \bar{x}(t+1) - x^\star \rangle + \|\eta_t \bar{g}(t)\|^2\end{aligned}$$

Thus,

$$\begin{aligned}\frac{1}{2}\|\bar{x}(t+1) - x^\star\|^2 &= \frac{1}{2}\|\bar{x}(t) - x^\star\|^2 \\ &\qquad - 2\eta_t \langle \bar{g}(t), \bar{x}(t+1) - x^\star \rangle - \eta_t^2 \|\bar{g}(t)\|^2 \\ &= \frac{1}{2}\|\bar{x}(t) - x^\star\|^2 + \eta_t \langle \bar{g}(t), x^\star - \bar{x}(t+1) \rangle \\ &\qquad - \frac{1}{2}\|\bar{x}(t+1) - \bar{x}(t)\|^2.\end{aligned}$$

Proof of Convergence II

We first bound the inner product term. We have

$$\langle \bar{g}(t), x^\star - \bar{x}(t+1)\rangle = \langle \bar{g}(t), x^\star - \bar{x}(t)\rangle + \langle \bar{g}(t), \bar{x}(t) - x(t+1)\rangle.$$

For the first term, we can apply the definition of $\bar{g}(t)$ and the convexity of each f_i to obtain

$$\begin{aligned}\langle \bar{g}(t), x^\star - \bar{x}(t)\rangle &= \frac{1}{N}\sum_{i=1}^{N}\langle \nabla f_i(x_i(t)), x^\star - \bar{x}(t)\rangle \\ &= \frac{1}{N}\sum_{i=1}^{N}\langle \nabla f_i(x_i(t)), x^\star - x_i(t)\rangle \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\langle \nabla f_i(x_i(t)), x_i(t) - \bar{x}(t)\rangle\end{aligned}$$

Proof of Convergence III

Continuing,

$$\begin{aligned}
\langle \bar{g}(t), x^\star - \bar{x}(t) \rangle &= \frac{1}{N}\sum_{i=1}^{N} \langle \nabla f_i(x_i(t)), x^\star - x_i(t) \rangle \\
&\quad + \frac{1}{N}\sum_{i=1}^{N} \langle \nabla f_i(x_i(t)), x_i(t) - \bar{x}(t) \rangle \\
&\leq \frac{1}{N}\sum_{i=1}^{N} (f_i(x^\star) - f_i(x_i(t))) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N} \langle \nabla f_i(x_i(t)), x_i(t) - \bar{x}(t) \rangle
\end{aligned}$$

Proof of Convergence IV

Thus,

$$\begin{aligned}\langle \bar{g}(t), x^\star - \bar{x}(t)\rangle &\le \frac{1}{N}\sum_{i=1}^{N}(f_i(x^\star) - f_i(x_i(t))) \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\langle \nabla f_i(x_i(t)), x_i(t) - \bar{x}(t)\rangle \\ &\le f(x^\star) - \hat{f}(X(t)),\end{aligned}$$

where

$$\begin{aligned}\hat{f}(X) &= \frac{1}{N}\sum_{i=1}^{N}(f_i(x_i + \langle \nabla f_i(x_i), \bar{x} - x_i\rangle) \\ &= \frac{1}{N}\left(F(X) + \mathrm{tr}\left(\nabla F(X)\left(\frac{1}{N}\mathbf{1}\mathbf{1}^\top X - X\right)^\top\right)\right).\end{aligned}$$

Proof of Convergence V

For the second term, we use the L-smoothness of each f_i to obtain

$$\begin{aligned} f_i(\bar{x}(t+1)) &\leq f_i(x_i(t)) + \langle \nabla f_i(x_i(t)), \bar{x}(t+1) - x_i(t) \rangle \\ &\qquad + \frac{L}{2} \|\bar{x}(t+1) - x_i(t)\|^2 \\ &= f_i(x_i(t)) + \langle \nabla f_i(x_i(t)), \bar{x}(t+1) - \bar{x}(t) \rangle \\ &\qquad + \langle \nabla f_i(x_i(t)), \bar{x}(t) - x_i(t) \rangle + \frac{L}{2} \|\bar{x}(t+1) - x_i(t)\|^2. \end{aligned}$$

Proof of Convergence VI

Thus,

$$\frac{1}{N}\sum_{i=1}^{N} f_i(\bar{x}(t+1)) \leq \frac{1}{N}\sum_{i=1}^{N}\Big(f_i(x_i(t)) + \langle \nabla f_i(x_i(t)), \bar{x}(t+1) - \bar{x}(t)\rangle$$

$$+ \langle \nabla f_i(x_i(t)), \bar{x}(t) - x_i(t)\rangle$$

$$+ \frac{L}{2}\|\bar{x}(t+1) - x_i(t)\|^2\Big)$$

$$= \hat{f}(X(t)) + \langle \bar{g}(t), \bar{x}(t+1) - \bar{x}(t)\rangle$$

$$+ \frac{L}{2N}\sum_{i=1}^{N}\|\bar{x}(t+1) - x_i(t)\|^2$$

Proof of Convergence VII

Using the fact that $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$.

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} f_i(\bar{x}(t+1)) &\leq \hat{f}(X(t)) + \langle \bar{g}(t), \bar{x}(t+1) - \bar{x}(t) \rangle \\
&\quad + \frac{L}{2N}\sum_{i=1}^{N} \|\bar{x}(t+1) - x_i(t)\|^2 \\
&\leq \hat{f}(X(t)) + \langle \bar{g}(t), \bar{x}(t+1) - \bar{x}(t) \rangle \\
&\quad + L\|\bar{x}(t+1) - \bar{x}(t)\|^2 + \frac{L}{N}\sum_{i=1}^{N} \|\bar{x}(t) - x_i(t)\|^2 \\
&= \hat{f}(X(t)) + \langle \bar{g}(t), \bar{x}(t+1) - \bar{x}(t) \rangle \\
&\quad + L\|\bar{x}(t+1) - \bar{x}(t)\|^2 + \frac{L}{N}\|E(t)\|_F^2.
\end{aligned}$$

Proof of Convergence VIII

Thus, combining the two bounds gives

$$\begin{aligned}
\langle \bar{g}(t), x^\star - \bar{x}(t+1)\rangle &= \langle \bar{g}(t), x^\star - \bar{x}(t)\rangle + \langle \bar{g}(t), \bar{x}(t) - x(t+1)\rangle \\
&\leq \left(f(x^\star) - \hat{f}(X(t)) \right) \\
&\quad + \Bigg(\hat{f}(X(t)) - f(\bar{x}(t+1)) + L\|\bar{x}(t+1) \\
&\quad\quad - \bar{x}(t)\|^2 + \frac{L}{N}\|E(t)\|_F^2 \Bigg) \\
&= f(x^\star) - f(\bar{x}(t+1)) + L\|\bar{x}(t+1) - \bar{x}(t)\|^2 \\
&\quad + \frac{L}{N}\|E(t)\|_F^2.
\end{aligned}$$

Proof of Convergence IX

Substitution into the original equality gives

$$\eta_t \left(f(\bar{x}(t+1) - f(x^\star) \right) \leq \frac{1}{2}\|\bar{x}(t) - x^\star\|^2 - \frac{1}{2}\|\bar{x}(t+1) - x^\star\|^2 - \frac{1}{2}\|\bar{x}(t+1) - \bar{x}(t)\|^2 + \eta_t L \|\bar{x}(t+1) - \bar{x}(t)\|^2 + \frac{\eta_t L}{N}\|E(t)\|_F^2.$$

Convergence Theorem for DGD

Theorem ([2, Lemma 3.2])

Suppose that $f_1, \ldots, f_N$ *are all convex and L-smooth, and there exists* $x^\star \in \mathbb{R}^d$ *such that* $x^\star \in \operatorname{argmin}_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^N f_i(x)$. *Consider the DGD algorithm Equation* (2.1) *such that Equation* (2.2) *holds and* $\eta_t > 0$ *for all* t. *Then, we have*

$$\frac{\sum_{\tau=1}^{t} \eta_{\tau-1}\left(f(\bar{x}(\tau)) - f(x^\star)\right)}{\sum_{\tau=1}^{t}\eta_{\tau-1}} \leq \frac{\|\bar{x}(0) - x^\star\|^2}{2\sum_{\tau=0}^{t-1}\eta_\tau} + \frac{L\sum_{\tau=0}^{t-1}\eta_\tau\|E(\tau)\|_F^2}{N\sum_{\tau=0}^{t-1}\eta_\tau} + \frac{\sum_{\tau=0}^{t-1}\eta_t^2(2\eta_\tau L - 1)\|\bar{g}(t)\|_F^2}{2\sum_{\tau=0}^{t-1}\eta_\tau}.$$

Constant Step Size Convergence Rate

Theorem

For simplicity, suppose every agent starts from the same initial point so that $E_0 = 0$. Suppose we choose a constant step size $\eta_t = \eta \leq 1/L$. Then

$$\frac{1}{t}\sum_{\tau=1}^{t}\left(f(\bar{x}(\tau)) - f(x^\star)\right) \leq \frac{\|\bar{x}(0) - x^\star\|^2}{2\eta t} + \frac{2\eta^2 L G^2}{(1-\sigma^2)^2},$$

and

$$\frac{1}{N}\sum_{i=1}^{N}\|x_i(t) - \bar{x}(t)\|^2 \leq \frac{\eta^2 G^2}{(1-\sigma^2)^2}.$$

Diminishing Step Size Convergence Rate

Theorem

Suppose we choose the step sizes to be $\eta_t = \frac{\alpha}{L(t+1)^\beta}$ *for some* $\alpha \in (0,1)$ *and* $\beta \in (0,1)$. *Then*

$$\frac{\sum_{\tau=1}^{t} \eta_{\tau-1}\left(f(\bar{x}(\tau)) - f(x^\star)\right)}{\sum_{\tau=1}^{t} \eta_{\tau-1}} \leq \begin{cases} O\left(\frac{1}{t^{2\beta}}\right), & 0 < \beta < 1/3, \\ O\left(\frac{\ln t}{t^{2/3}}\right), & \beta = 1/3, \\ O\left(\frac{1}{t^{1-\beta}}\right), & 1/3 < \beta < 1, \end{cases}$$

and

$$\frac{1}{N}\sum_{i=1}^{N} \|x_i(t) - \bar{x}(t)\|^2 \leq O\left(\frac{1}{t^{2\beta}}\right).$$

Discussion

- A natural question is if the Lipschitz continuity of the local cost functions $f_1, \ldots, f_N$ can be relaxed. [1] shows that for DGD (and diffusion) the answer is no:

Counterexample

Let $f_i^\theta(x) = \frac{1}{2}(x + (-1)^i\theta)^2$, so each f_i^θ is convex and 1-smooth, but not Lipschitz. Consider the weight matrix $W = \frac{1}{4}\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$, with initial points $x_1(0) = x_2(0) = 0$. Then, we can show that
$x_1(t) = -x_2(t) \geq 0$ and

$$x_1(t+1) = \frac{3}{4}x_1(t) - \frac{1}{4}x_2(t) - \eta_t(x_1(t) - \theta) = \left(\frac{1}{2} - \eta_t\right) x_1(t)\eta_t\theta$$

Thus,

$$\frac{1}{2}\left(\|x_1(t) - \bar{x}(t)\|^2 + \|x_2(t) - \bar{x}(t)\|^2 +\right) \geq \eta_t\theta^2 = \frac{\eta_0\theta^2}{(t+1)^{2\beta}}.$$

Discussion

- The previous example shows that Lipschitz continuity of the local cost functions $f_1, \ldots, f_N$ is needed to ensure that the consensus error $\|E(t)\|_F$ decays fast enough.

Discussion

- The previous example shows that Lipschitz continuity of the local cost functions $f_1, \dots, f_N$ is needed to ensure that the consensus error $\|E(t)\|_F$ decays fast enough.
- Also, in general a constant stepsize $\eta_t = \eta$ converges to a neighborhood within $O\left(\frac{\eta^2}{(1-\sigma)^2}\right)$ of the optimal cost

Discussion

- The previous example shows that Lipschitz continuity of the local cost functions $f_1, \ldots, f_N$ is needed to ensure that the consensus error $\|E(t)\|_F$ decays fast enough.
- Also, in general a constant stepsize $\eta_t = \eta$ converges to a neighborhood within $O\left(\frac{\eta^2}{(1-\sigma)^2}\right)$ of the optimal cost
  - This is partly because the optimal solution is not a fixed point of the DGD update

Discussion

- The previous example shows that Lipschitz continuity of the local cost functions $f_1, \ldots, f_N$ is needed to ensure that the consensus error $\|E(t)\|_F$ decays fast enough.
- Also, in general a constant stepsize $\eta_t = \eta$ converges to a neighborhood within $O\left(\frac{\eta^2}{(1-\sigma)^2}\right)$ of the optimal cost
  - This is partly because the optimal solution is not a fixed point of the DGD update
- A diminishing sequence of stepsizes does converge, but at an inferior rate.

Discussion

- The previous example shows that Lipschitz continuity of the local cost functions $f_1, \dots, f_N$ is needed to ensure that the consensus error $\|E(t)\|_F$ decays fast enough.
- Also, in general a constant stepsize $\eta_t = \eta$ converges to a neighborhood within $O\left(\frac{\eta^2}{(1-\sigma)^2}\right)$ of the optimal cost
  - This is partly because the optimal solution is not a fixed point of the DGD update
- A diminishing sequence of stepsizes does converge, but at an inferior rate.
  - The best stepsize, $\beta = \frac{1}{3}$, gives an $O\left(t^{-2/3} \ln t\right)$ rate, slower than $O\left(t^{-1}\right)$ in the centralized case
- [1] showed an $\Omega(t^{-2/3})$ lower bound on the worst-case performance

Motivation

- The main issue with the previous methods is the inexactness of the gradient update.
- If each agent knew the true gradient of $\bar{x}(t)$, then we could recover the convergence rate of the centralized case.
- Some issues
  - How do we estimate the gradient more accurately?
  - The $O(t^{-1})$ convergence rate holds for constant stepsizes, but then the consensus error may not decay fast enough!
- Solution:
  - Add an auxiliary variable to estimate the global gradient
  - Add extra update step to ensure consensus error decays fast enough

Gradient Tracking

Definition

The **gradient tracking algorithm** is

$$x_i(t+1) = \sum_{j=1}^{N} W_{ij} x_j(t) - \eta g_i(t)$$

$$g_i(t+1) = \sum_{j=1}^{N} W_{ij} g_j(t) - \nabla f_i(x_i(t+1)) - \nabla f_i(x_i(t)).$$

Gradient Tracking

Definition

The **gradient tracking algorithm** is

$$x_i(t+1) = \sum_{j=1}^{N} W_{ij} x_j(t) - \eta g_i(t)$$

$$g_i(t+1) = \sum_{j=1}^{N} W_{ij} g_j(t) - \nabla f_i(x_i(t+1)) - \nabla f_i(x_i(t)).$$

Note that each agent needs to communicate its state variable $x_i(t)$ and the gradient estimate $g_i(t)$ at each time step.

Notation

We introduce the notations

$$\bar{x}(t) := \frac{1}{N}\sum_{i=1}^{N} x_i(t), \quad \bar{g}(t) := \frac{1}{N}\sum_{i=1}^{N} g_i(t).$$

We also use $E_x(t)$ and $E_g(t)$ to denote the consensus errors:

$$E_x(t) := \begin{bmatrix} (x_1(t) - \bar{x}(t))^\top \\ \vdots \\ (x_N(t) - \bar{x}(t))^\top \end{bmatrix} = \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \right) X(t),$$

$$E_g(t) := \begin{bmatrix} (g_1(t) - \bar{g}(t))^\top \\ \vdots \\ (g_N(t) - \bar{g}(t))^\top \end{bmatrix} = \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \right) G(t).$$

Proof Sketch I

The proof is very lengthy and technical, but we cover some highlights:

(1) Show that

$$E_x(t) = \left(I - \mathbf{1}\mathbf{1}^\top\right) X(t)$$
$$E_x(t) = \left(I - \mathbf{1}\mathbf{1}^\top\right) G(t)$$
$$\bar{x}(t+1) = \bar{x}(t) - \eta_t \bar{g}(t)$$
$$\|\bar{g}(t) - \nabla f(\bar{x}(t))\| \leq \frac{L}{\sqrt{N}} \|E_x(t)\|_F$$

Proof Sketch II

(2) Using (1), show that

$$\begin{bmatrix} \|E_x(t+1)\|_F \\ \frac{\eta}{2L\sqrt{3}}\|E_g(t+1)\|_F \end{bmatrix} \leq \overbrace{\begin{bmatrix} \frac{1+\sigma}{2} & \frac{2\eta L\sqrt{3}}{1-\alpha} \\ \frac{2\eta L\sqrt{3}}{1-\alpha} & \frac{1+\sigma}{2} \end{bmatrix}}^{P(\eta L)} \begin{bmatrix} \|E_x(t+1)\|_F \\ \frac{\eta}{2L\sqrt{3}}\|E_g(t+1)\|_F \end{bmatrix} + \frac{\eta^2 LN\sqrt{3}}{2(1-\sigma)} \begin{bmatrix} 0 \\ \|\bar{g}(t)\|^2 \end{bmatrix}$$

(3) Show that $\|P(\varepsilon)\| \leq \frac{2+\sigma}{3}$ for $\varepsilon \in \left(0, \frac{(1-\sigma)^2}{25}\right)$.

Proof Sketch III

(4) Conclude that

$$\sum_{\tau=0}^{t-1} \|E_x(\tau)\|_F^2 \leq \frac{2NE_0^2}{1-\sigma} + \frac{3\eta^3 LN\sqrt{3}}{2(1-\sigma)^2} \sum_{\tau=0}^{t-1} \|\bar{g}(t)\|^2$$

where E_0 is some constant depending only on the initial conditions.

(5) Use the theorem from earlier to conclude that

$$\frac{1}{t} \sum_{\tau=1}^{t} (f(\bar{x}(\tau)) - f(x^\star)) \leq \frac{1}{t} \left[\frac{\|\bar{x}(0) - x^\star\|^2}{2\eta} + \frac{3LE_0^2}{2(1-\sigma)} \right]$$

Proof Sketch IV

(6) Show that

$$\sum_{\tau=0}^{t-1} \|\bar{g}(t)\|^2 \leq \frac{5}{2\eta} \left[\frac{\|\bar{x}(0) - x^\star\|^2}{2\eta} + \frac{3LE_0^2}{2(1-\sigma)} \right]$$

(6.5) We need to use the general fact that if $\sum_{n=0}^{\infty} a_n < \infty$ then $\min_{0 \leq n \leq t-1} a_n = o(t^{-1})$.

(7) Conclude using (6) and (6.5) that

$$\min_{\tau=0,\ldots,t-1} \|E_x(\tau)\|_F^2 \leq o(t^{-1})$$

Theorem

For gradient tracking, we have

$$\frac{1}{t}\sum_{\tau=1}^{t}\left(f(\bar{x}(\tau)) - f(x^\star)\right) \leq \frac{1}{t}\left[\frac{\|\bar{x}(0) - x^\star\|^2}{2\eta} + \frac{3LE_0^2}{2(1-\sigma)}\right]$$

$$\min_{\tau=0,\dots,t-1}\|E_x(\tau)\|_F^2 \leq o(t^{-1})$$

Theorem

For gradient tracking, we have

$$\frac{1}{t}\sum_{\tau=1}^{t}\left(f(\bar{x}(\tau)) - f(x^\star)\right) \leq \frac{1}{t}\left[\frac{\|\bar{x}(0) - x^\star\|^2}{2\eta} + \frac{3LE_0^2}{2(1-\sigma)}\right]$$

$$\min_{\tau=0,\dots,t-1} \|E_x(\tau)\|_F^2 \leq o(t^{-1})$$

- Gradient tracking recovers the centralized convergence rate in the smooth convex case

Theorem

For gradient tracking, we have

$$\frac{1}{t}\sum_{\tau=1}^{t}(f(\bar{x}(\tau)) - f(x^\star)) \leq \frac{1}{t}\left[\frac{\|\bar{x}(0) - x^\star\|^2}{2\eta} + \frac{3LE_0^2}{2(1-\sigma)}\right]$$

$$\min_{\tau=0,\dots,t-1}\|E_x(\tau)\|_F^2 \leq o(t^{-1})$$

- Gradient tracking recovers the centralized convergence rate in the smooth convex case
-
    - Similar results hold in the strongly convex case

Theorem

For gradient tracking, we have

$$\frac{1}{t}\sum_{\tau=1}^{t}\left(f(\bar{x}(\tau)) - f(x^\star)\right) \leq \frac{1}{t}\left[\frac{\|\bar{x}(0) - x^\star\|^2}{2\eta} + \frac{3LE_0^2}{2(1-\sigma)}\right]$$

$$\min_{\tau=0,\ldots,t-1} \|E_x(\tau)\|_F^2 \leq o(t^{-1})$$

- Gradient tracking recovers the centralized convergence rate in the smooth convex case
- - Similar results hold in the strongly convex case
- Notice the sensitivity to the parameter σ, if σ is close to 1 then the problem is poorly conditioned and gradient tracking may still perform poorly

Theorem

For gradient tracking, we have

$$\frac{1}{t}\sum_{\tau=1}^{t}(f(\bar{x}(\tau)) - f(x^\star)) \leq \frac{1}{t}\left[\frac{\|\bar{x}(0) - x^\star\|^2}{2\eta} + \frac{3LE_0^2}{2(1-\sigma)}\right]$$

$$\min_{\tau=0,\ldots,t-1} \|E_x(\tau)\|_F^2 \leq o(t^{-1})$$

- Gradient tracking recovers the centralized convergence rate in the smooth convex case
  - Similar results hold in the strongly convex case
- Notice the sensitivity to the parameter σ, if σ is close to 1 then the problem is poorly conditioned and gradient tracking may still perform poorly
- The technique of constructing an associated linear dynamical system is somewhat common in the literature

Conclusion

- We have introduced two algorithms for smooth distributed convex optimization
- We discussed the features of each, as well as their convergence rates
- We proved the convergence of one naive algorithm for distributed optimization, as well as discussed its limitations.
- There are much more advanced and preferable algorithms, some of which incorporate addition internal dynamics to offset the negative effects of consensus error on the gradient updates.
- A key takeaway is that distributed first-order algorithms are theoretically similar to inexact first-order methods, where controlling the inexactness is needed to ensuring convergence.

References I

[1] Jakovetić, D., Xavier, J., and Moura, J. M. F. (2014). Fast distributed gradient methods. *IEEE Transactions on Automatic Control*, 59(5):1131–1146.

[2] Tang, Y. (2024). Fundamentals of distributed optimization.