Fundamentals of Distributed Optimization

Lecture Notes

Yujie Tang

Spring, 2024

WORKING DRAFT:

Please email yujietang@pku.edu.cn if you find any errors or typos or have any comments. Your feedback would be greatly appreciated!

Contents

Chapter 1

Preliminaries

1.1 Some Analysis and Linear Algebra

Notations. In this course, we exclusively consider *finite-dimensional* optimization problems defined over standard (real) Euclidean spaces. The standard inner product will be denoted by $\langle x, y\rangle \coloneqq x^\mathsf{T} y$ for $x, y \in \mathbb{R}^n$, and the induced norm (the ℓ_2 norm) will be denoted by $\|x\| \coloneqq \sqrt{\langle x, x\rangle}$. The vector in $\mathbb{R}^n$ will all entries equal to 1 will be denoted by $\mathbf{1}$.

Basic Notions and Facts in Analysis

- Let S be a subset of $\mathbb{R}$. We say that $c \in \mathbb{R}$ is a lower bound of S, if $x \geq c$ for all $x \in S$. Similarly, we can define the notion of upper bounds for a subset of $\mathbb{R}$.

  It is a classical result in analysis that when $S \subseteq \mathbb{R}$ admits a lower bound, there is a lower bound of S that is greater than or equal to any lower bound of S. This greatest lower bound of S will be called the *infimum* of S and will be denoted by $\inf S$.

  Similarly, when S admits an upper bound, the least upper bound of S will be called the *supremum* of S and will be denoted by $\sup S$.

- A sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{R}^n$ is said to *converge* to $x \in \mathbb{R}^n$, if for any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $\|x_n - x\| < \epsilon$ whenever $n > N$.

  Equivalently, we say that x_n converges to x if and only if $\lim_{n\to\infty} \|x_n - x\| = 0$.

- A set $O \subseteq \mathbb{R}^n$ is said to be *open*, if for any $x \in O$, there exists $\epsilon > 0$ such that $\{y \in \mathbb{R}^n : \|y - x\| < \epsilon\} \subseteq O$, i.e., O contains a ball of radius ϵ centered around x.

  It can be shown that the union of an arbitrary family of open sets is open.

- The *interior* of a set $S \subseteq \mathbb{R}^n$, denoted by $\operatorname{int} S$, is the union of all open sets that are subsets of S. Equivalently, $x \in \operatorname{int} S$ if and only if there exists $\epsilon > 0$ such that $\{y \in \mathbb{R}^n : \|y - x\| < \epsilon\} \subseteq S$.

- A set $C \subseteq \mathbb{R}^n$ is said to be *closed* if for any sequence $(x_n)_{n\in\mathbb{N}}$ in C, whenever x_n converges to some $x \in \mathbb{R}^n$, we have $x \in C$.

  It can be shown that a set $C \subseteq \mathbb{R}^n$ is closed if and only if its complement $\mathbb{R}^n \backslash C$ is open.

  It can be shown that the intersection of an arbitrary family of closed sets is closed.

- The *closure* of a set $S \subseteq \mathbb{R}^n$, denoted by $\operatorname{cl} S$, is the intersection of all closed subsets of $\mathbb{R}^n$ that contain S as a subset. Equivalently, $x \in \operatorname{cl} S$ if and only if there exists a sequence $(x_n)_{n\in\mathbb{N}}$ in S such that $x_n \to x$.

- The *boundary* of a set $S \subseteq \mathbb{R}^n$, denoted by ∂S, is defined by $\partial S \coloneqq \operatorname{cl} S \backslash \operatorname{int} S$. Equivalently, $x \in \partial S$ if and only if for any $\epsilon > 0$, the set $\{y \in \mathbb{R}^n : \|y - x\| < \epsilon\}$ intersects with both S and $\mathbb{R}^n \backslash S$.

- A sequence $(x_n)_{n\in\mathbb{N}}$ is said to be *Cauchy*, if for any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $\|x_m - x_n\| < \epsilon$ whenever $m > N$ and $n > N$.

  Any closed subset $C \subseteq \mathbb{R}^n$ is *complete* in the sense that any Cauchy sequence (x_n) in C converges to some $x \in C$.

- An open cover of a set $S \subseteq \mathbb{R}^n$ is a family of open sets $\{O_\alpha \subseteq \mathbb{R}^n : \alpha \in \mathcal{I}\}$ such that $S \subseteq \bigcup_{\alpha\in\mathcal{I}} O_\alpha$. Note that the index set $\mathcal{I}$ can be arbitrary (e.g., uncountable).

  A set $C \subseteq \mathbb{R}^n$ is said to be *compact*, if for any open cover $\{O_\alpha : \alpha \in \mathcal{I}\}$ of S, there exists a finite subset $\mathcal{J} \subseteq \mathcal{I}$ such that $S \subseteq \bigcup_{\alpha\in\mathcal{J}} O_\alpha$.

  If a set $C \subseteq \mathbb{R}^n$ is compact, then for any sequence $(x_n)_{n\in\mathbb{N}}$ in C, there exists a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ that converges to some $x \in C$.

  By the Heine-Borel theorem, a set $C \subseteq \mathbb{R}^n$ is compact if and only if C is closed and there exists $R > 0$ such that $C \subseteq \{x \in \mathbb{R}^n : \|x\| \leq R\}$.

- Let $f : S \to \mathbb{R}$ where $S \subseteq \mathbb{R}^n$. Given $x \in S$, f is said to be *continuous* at x if for any $\epsilon > 0$, there exists $\delta > 0$ such that $|f(y) - f(x)| < \epsilon$ whenever $y \in S$ and $\|y - x\| < \delta$. f is said to be a *continuous function* (or *continuous over* S, or simply *continuous*), if f is continuous at every point in S.

  By the extreme value theorem, if $C \subseteq \mathbb{R}^n$ is a compact set and $f : C \to \mathbb{R}$ is a continuous function, then f attains its maximum value and minimum value on C.

  We refer to [Rudin, 1976] for more details and proofs of these results.

Basic Notions and Facts in Linear Algebra

Eigenvalues and Spectral Decomposition

- For a complex squared matrix $A \in \mathbb{C}^{n\times n}$, whenever $Ax = \lambda x$ for some $\lambda \in \mathbb{C}$ and $x \in \mathbb{C}^n \backslash \{0\}$, we say that λ is an *eigenvalue* of A and x is the associated *eigenvector*.

- The Spectral Theorem: Let $A \in \mathbb{C}^{n\times n}$, and let $\lambda_1, \ldots, \lambda_p$ be the distinct eigenvalues of A. Denote

$$\operatorname{null}(A - \lambda_i I)^n = \{v \in \mathbb{C}^c : (A - \lambda_i I)^n v = 0\},$$

i.e., $\text{null}(A-\lambda_i I)^n$ is the null space of the linear operator $(A-\lambda_i I)^n$. Then any $x \in \mathbb{C}^n$ can be uniquely decomposed as

$$x = x_1 + x_2 + \cdots + x_n$$

with $x_i \in \text{null}(A-\lambda_i I)^n$ for each $i = 1, \ldots, p$.

- Jordan canonical form: For $\lambda \in \mathbb{C}$ and positive integer n, we introduce the notation

$$J_1(\lambda) = [\lambda], \qquad J_2(\lambda) = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \qquad J_n(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & \cdots & 0 & \lambda \end{bmatrix} \in \mathbb{C}^{n\times n}.$$

  Let $A \in \mathbb{C}^{n\times n}$ be arbitrary. Then there exist an invertible matrix $S \in \mathbb{C}^{n\times n}$ and a matrix $J \in \mathbb{C}^{n\times n}$ such that

  1. $S^{-1}AS = J$.
  2. The matrix J is a block-diagonal matrix of the form

$$J = \begin{bmatrix} J_{n_1}(\lambda_1) & & & \\ & J_{n_2}(\lambda_2) & & \\ & & \ddots & \\ & & & J_{n_q}(\lambda_q) \end{bmatrix}$$

     for some $\lambda_1, \ldots, \lambda_q \in \mathbb{C}$ and positive integers $n_1, \ldots, n_q$ with $n_1 + \cdots + n_q = n$; $\lambda_1, \ldots, \lambda_q$ are not necessarily distinct.

  The matrix J is called the *Jordan canonical form* of A.

- An eigenvalue of $A \in \mathbb{C}^{n\times n}$ is called *simple*, if $\dim \text{null}(A-\lambda I)^n = 1$. It can be shown that the following statements are equivalent:

  1. λ is a simple eigenvalue of A.
  2. λ appears exact once in the diagonal of the Jordan canonical form of A.
  3. $\text{null}(A-\lambda I) = \text{span}\{v\}$ for some $v \in \mathbb{C}^n \backslash \{0\}$, and there exists no vector u such that $Au = \lambda u + v$.

Real Symmetric Matrices

- We say that $A \in \mathbb{R}^{n\times n}$ is *real symmetric* if $A = A^\mathsf{T}$. It is a standard result in linear algebra that for a real symmetric matrix A, there exists $U \in \mathbb{R}^{n\times n}$ satisfying $U^\mathsf{T}U = I$ such that

$$A = U\Lambda U^\mathsf{T},$$

  which can also be equivalently written as

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\mathsf{T}.$$

Here $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and u_i is the i'th column vector of U. It can be seen that each u_i is an eigenvector of A with eigenvalue λ_i satisfying $\|u_i\| = 1$.

- We say that a real symmetric matrix A is *positive semidefinite* if $x^\mathsf{T} A x \geq 0$ for all x. It is called *positive definite* if $x^\mathsf{T} A x > 0$ as long as $x \neq 0$. A real symmetric matrix A is positive semidefinite (resp. positive definite) if and only if all eigenvalues of A are nonnegative (resp. positive). For real symmetric matrices A, B, we denote $A \succeq B$ or $B \preceq A$ if $A - B$ is positive semidefinite, and denote $A \succ B$ and $B \prec A$ if $A - B$ is positive definite.

Matrix Norms and Spectral Radius

- For a real matrix A, we define its *spectral norm* by

$$\|A\|_2 \coloneqq \max_{\|x\|=1} \|Ax\|.$$

It can be shown that $\|A\|_2$ is equal to the largest singular value of A, or equivalently, $\|A\|_2^2$ is equal to the largest eigenvalue of $A^\mathsf{T} A$ (and AA^T). By definition, we have

$$\|Ax\| \leq \|A\|_2 \cdot \|x\|$$

for all x.

- The *Frobenius norm* of a real matrix A is defined by

$$\|A\|_F \coloneqq \sqrt{\mathrm{tr}(A^\mathsf{T} A)} = \sqrt{\sum\nolimits_{i,j} |A_{ij}|^2}.$$

- Now suppose A is a square matrix, its spectral radius, denoted by $\rho(A)$, is defined as the maximum magnitude of its eigenvalues. Note that $\rho(A)$ is in general not a norm.

We have the following facts about the spectral radius of $A \in \mathbb{R}^{n \times n}$:

1. $\lim_{k \to \infty} A^k = 0$ if and only if $\rho(A) < 1$.
2. $\rho(A) \leq \left\|A^k\right\|_2^{1/k}$ for all $k \geq 1$.
3. (Gelfand's formula) $\rho(A) = \lim_{k \to \infty} \left\|A^k\right\|_2^{1/k}$.

We refer to [Lax, 2007] and [Horn and Johnson, 2013] for more details and proofs of these results.

Exercise 1.1. Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$. Show that

$$\|AB\|_F \leq \|A\|_2 \cdot \|B\|_F. \qquad \square$$

1.2 Convex Sets and Functions

Definition 1.1. A set $C \subseteq \mathbb{R}^n$ is called *convex*, if for any $x, y \in C$ and any $\alpha \in [0, 1]$, we have $\alpha x + (1 - \alpha) y \in C$. Geometrically, this means that any line segments connecting two points x, y in a convex set C will always be contained in the set C.

Theorem 1.1 (Projection onto closed convex sets)**.** *Let $C \subseteq \mathbb{R}^n$ be a closed convex set. Then for any $x \in \mathbb{R}^n$, there exists a unique $z^* \in C$ such that $\|x - z^*\| \leq \|x - z\|$ for any $z \in C$.*

The point z^* will be called the projection of x onto C, and will be denoted by $\mathcal{P}_C[x]$.

Proof. We first show the existence of z^*. Fix $x \in \mathbb{R}^n$, and denote $\delta \coloneqq \inf\{\|z - x\| : z \in C\}$. It is evident that $\delta \geq 0$. Now let $(z_k)_{k\geq 1}$ be a sequence of points in C such that

$$\|z_k - x\|^2 \leq \delta^2 + \frac{1}{k}$$

for any k. We first show that (z_k) is a Cauchy sequence. Let $k, l > 0$ be arbitrary, and notice that, since C is convex, we have $\frac{1}{2}(z_k + z_l) \in C$, which implies

$$\left\|\frac{1}{2}(z_k + z_l) - x\right\|^2 \geq \delta^2.$$

Expanding this inequality leads to

$$\frac{1}{2}\langle z_k - x, z_l - x\rangle \geq \delta^2 - \frac{1}{4}\|z_k - x\|^2 - \frac{1}{4}\|z_l - x\|^2. \tag{1.1}$$

We now calculate $\|z_k - z_l\|^2$ and get

$$\begin{aligned}\|z_k - z_l\|^2 &= \|z_k - x\|^2 + \|z_l - x\|^2 - 2\langle z_k - x, z_l - x\rangle \\ &\leq 2(\|z_k - x\|^2 + \|z_l - x\|^2) - 4\delta^2 \leq \frac{2}{k} + \frac{2}{l}.\end{aligned}$$

where we used the inequality (1.1) in the second step. Thus for any $\epsilon > 0$, as long as $k, l \geq \lceil 4/\epsilon^2 \rceil$, we have $\|z_k - z_l\| \leq \epsilon$, showing that (z_k) is a Cauchy sequence. By the completeness of $\mathbb{R}^d$, $z^* = \lim_{z\to\infty} z_k$ exists. Since C is closed, we have $z^* \in C$, and by the continuity of the norm function, we have

$$\|z^* - x\| = \lim_{k\to\infty} \|z_k - x\| = \delta,$$

which is less than or equal to $\|z - x\|$ for any $z \in C$, by the definition of δ.

Next, we show the uniqueness of z^*. Suppose both $u^{*(1)}$ and $z^{*(2)}$ satisfies $\left\|z^{*(1)} - x\right\| = \left\|z^{*(2)} - x\right\| = \delta$. Denote $\tilde{z} = \frac{1}{2}(z^{*(1)} + z^{*(2)})$, and we have

$$\delta^2 \leq \|\tilde{z} - x\|^2 = \left\|\frac{1}{2}(z^{*(1)} - x) + \frac{1}{2}(z^{*(2)} - x)\right\|^2 = \frac{1}{2}\delta^2 + \frac{1}{2}\langle z^{*(1)} - x, z^{*(2)} - x\rangle,$$

which leads to

$$\langle z^{*(1)} - x, z^{*(2)} - x\rangle \geq \delta^2.$$

Consequently,

$$\left\|z^{*(1)} - z^{*(2)}\right\|^2 = \|z^{*(1)} - x\|^2 + \|z^{*(2)} - x\|^2 - 2\langle z^{*(1)} - x, z^{*(2)} - x\rangle \leq \delta^2 + \delta^2 - 2\delta^2 = 0,$$

implying that $z^{*(1)} = z^{*(2)}$. The proof is now complete. □

Exercise 1.2. Let $C \subseteq \mathbb{R}^n$ be a closed convex set. Prove that

1. For any $x \in \mathbb{R}^n$ and $y \in C$, we have $y = \mathcal{P}_C[x]$ if and only if $\langle z - y, x - y \rangle \leq 0$ for any $z \in C$.
2. (*Nonexpansiveness*) $\|\mathcal{P}_C[x] - \mathcal{P}_C[y]\| \leq \|x - y\|$ for any $x, y \in \mathbb{R}^n$.

   (Hint: Expand $\|x - y\|^2 = \|(x - \mathcal{P}_C[x] + \mathcal{P}_C[y] - y) + (\mathcal{P}_C[x] - \mathcal{P}_C[y])\|^2$ and use the previous result.) □

Theorem 1.2 (Supporting Hyperplane Theorem)**.** *Let $C \subseteq \mathbb{R}^n$ be a closed convex set, and let x be a boundary point of C. Then there exists a nonzero $v \in \mathbb{R}^n$ such that*

$$\langle v, z - x \rangle \geq 0$$

for all $z \in C$.

Proof. Since x is a boundary point of C, for each $k \geq 1$, we can find $x_k \in \mathbb{R}^n \backslash C$ such that $\|x_k - x\| \leq 1/k$. By the second part of Exercise 1.2, we have

$$\langle z - \mathcal{P}_C[x_k], x_k - \mathcal{P}_C[x_k] \rangle \leq 0 \tag{1.2}$$

for any $z \in C$ and any $k \geq 1$ by the convexity of C. Now define

$$v_k = \frac{\mathcal{P}_C[x_k] - x_k}{\|\mathcal{P}_C[x_k] - x_k\|}.$$

Obviously $\|v_k\| = 1$, and by the compactness of the unit sphere in $\mathbb{R}^n$, there exists a subsequence $(v_{k_m})_{m \geq 1}$ such that $v_{k_m} \to v$ as $m \to \infty$. On the other hand, by (1.2), for any $z \in C$,

$$\langle z - \mathcal{P}_C[x_{k_m}], v_{k_m} \rangle \geq 0.$$

By taking the limit $m \to \infty$ and noting that $\|\mathcal{P}_C[x_{k_m}] - x\| = \|\mathcal{P}_C[x_{k_m}] - \mathcal{P}_C[x]\| \leq \|x_{k_m} - x\| \to 0$, we obtain

$$\langle z - x, v \rangle \geq 0$$

for all $z \in C$, which completes the proof. □

Let $f : \mathcal{S} \to \mathbb{R}$ where $\mathcal{S} \subseteq \mathbb{R}^n$. The *epigraph* of f on a set $\mathcal{X} \subseteq \mathcal{S}$ is defined as

$$\operatorname{epi}(f; \mathcal{X}) \coloneqq \{(x, y) \in \mathcal{X} \times \mathbb{R} : f(x) \leq y\}.$$

Definition 1.2. A function $f : \mathcal{S} \to \mathbb{R}$ is said to be *convex* on a set $\mathcal{X} \subseteq \mathcal{S}$ if the epigraph $\operatorname{epi}(f; \mathcal{X})$ is a convex set. Equivalently, f is *convex* on $\mathcal{X}$ if and only if $\mathcal{X}$ is convex, and for any $x, y \in \mathcal{X}$ and $\alpha \in [0, 1]$, we have

$$f(\alpha x + (1 - \alpha) y) \leq \alpha f(x) + (1 - \alpha) f(y).$$

When f is convex on its domain $\mathcal{S}$, we simply say that f is convex.

Lemma 1.1 (Jensen's inequality)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be convex. Then for any $x_1, \ldots, x_m \in \mathcal{X}$ and any nonnegative $\alpha_1, \ldots, \alpha_m$ satisfying $\sum_i \alpha_i = 1$, we have*

$$f\left(\sum_{i=1}^m \alpha_i x_i\right) \leq \sum_{i=1}^m \alpha_i f(x_i).$$

Definition 1.3. A function $f : \mathcal{S} \to \mathbb{R}$ is called *μ-strongly convex* on $\mathcal{X} \subseteq \mathcal{S}$, if the function $x \mapsto f(x) - \frac{1}{2}\mu\|x\|^2$ is convex on $\mathcal{X}$.

Exercise 1.3. Let $f : [0, \delta] \to \mathbb{R}$ be a convex function for some $\delta > 0$. Define

$$h(t) = \frac{f(t) - f(0)}{t}, \qquad 0 < t \leq \delta$$

Show that $h(t)$ is a non-decreasing function over $t \in (0, \delta]$.

(Hint: Notice that for $0 < t_1 < t_2 \leq \delta$, we have $t_1 = (t_1/t_2) \cdot t_2 + (1 - t_1/t_2) \cdot 0$. We can then apply the definition of convexity to upper bound $f(t_1)$.) □

Exercise 1.4.

1. Suppose $f : (-\delta, \delta) \to \mathbb{R}$ is differentiable, and is convex on $[0, \delta)$. Show that
$$f(t) \geq f(0) + tf'(0), \qquad \forall t \in [0, \delta).$$
(Hint: Note that $(f(t) - f(0))/t$ is non-decreasing and converges to $f'(0)$ as $t \downarrow 0$.)
2. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, and is convex on $\mathcal{X} \subseteq \mathbb{R}^n$. Show that
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \qquad \forall x, y \in \mathcal{X}.$$ □

We shall provide a stronger version of the second part of Exercise 1.4.

Proposition 1.1. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, and $\mathcal{X} \subseteq \mathbb{R}^n$ is convex. Then f is convex on $\mathcal{X}$ if and only if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \qquad \forall x, y \in \mathcal{X} \tag{1.3}$$

if and only if

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0, \qquad \forall x, y \in \mathcal{X}. \tag{1.4}$$

Proof. The part that convexity implies (1.3) has been shown in the second part of Exercise 1.4. To show that (1.3) implies convexity, we note that 1.1 implies

$$f(x) \geq f(\alpha x + (1-\alpha)y) + \langle \nabla f(\alpha x + (1-\alpha)y), (1-\alpha)(x-y) \rangle,$$
$$f(y) \geq f(\alpha x + (1-\alpha)y) + \langle \nabla f(\alpha x + (1-\alpha)y), \alpha(y-x) \rangle$$

for any $x, y \in \mathcal{X}$ and $\alpha \in [0,1]$. By multiplying the first inequality by α, the second by $(1-\alpha)$ and summing them together, we get

$$\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha) y),$$

which justifies that f is convex on $\mathcal{X}$.

The inequality (1.4) follows by adding two copies of (1.3) with x and y interchanged. To show that (1.4) implies (1.3), we note that, by the Newton-Leibniz Theorem,

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle \nabla f(x + t(y-x)), y - x \rangle \, dt \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle \, dt \\ &\geq f(x) + \langle \nabla f(x), y - x \rangle, \end{aligned}$$

where in the last step we used $\langle \nabla f(x+t(y-x)) - \nabla f(x), t(y-x) \rangle \geq 0$ that follows from (1.4). □

Exercise 1.5. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, and $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set. Show that f is μ-strongly convex on $\mathcal{X}$ if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2, \qquad \forall x, y \in \mathcal{X}$$

if and only if

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \mu \|y - x\|^2, \qquad \forall x, y \in \mathcal{X}. \qquad \square$$

Definition 1.4. Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function. Given $x \in \mathcal{X}$, a vector $g \in \mathbb{R}^n$ is called a *subgradient* of f, if

$$f(y) \geq f(x) + \langle g, y - x \rangle$$

for all $y \in \mathcal{X}$. The set of all subgradients of f at x is called the *subdifferential* of f at x, and is denoted by $\partial f(x)$.

Lemma 1.2. *Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function, and let x be any point that lies in the interior of $\mathcal{X}$. Then the following statements hold:*

1. *$\partial f(x)$ is nonempty.*
2. *If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$. Conversely, if $\partial f(x)$ contains only one element, then f is differentiable at x.*

Proof (First part). Since f is convex, its epigraph $\mathrm{epi}(f)$ is a convex set. It's not hard to see that $(x, f(x))$ is a boundary point of $\mathrm{epi}(f)$, and so by the Supporting Hyperplane Theorem, there exists a nonzero $v = (\tilde{v}, v_{n+1}) \in \mathbb{R}^{n+1}$ such that

$$\begin{bmatrix} \tilde{v} \\ v_{n+1} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} x' - x \\ y' - f(x) \end{bmatrix} \geq 0, \qquad \forall (x', y') \in \mathrm{epi}(f),$$

i.e.,

$$\tilde{v}^{\mathsf{T}}(x' - x) + v_{n+1}(y' - f(x)) \geq 0, \qquad \forall x' \in \mathcal{X}, y' \geq f(x'). \tag{1.5}$$

It can be seen that $v_{n+1} \geq 0$; otherwise the above inequality would not hold for sufficiently large y'. Moreover, v_{n+1} cannot be zero: $v_{n+1} = 0$ would lead to

$$\tilde{v}^{\mathsf{T}}(x' - x) \geq 0, \qquad \forall x' \in \mathcal{X},$$

and since x is in the interior of $\mathcal{X}$, we can find a sufficiently small $\delta > 0$ such that $x - \delta\tilde{v} \in \mathcal{X}$. By plugging in $x' = x - \delta\tilde{v}$, we get

$$\|\tilde{v}\|^2 \leq 0,$$

implying that $\tilde{v} = 0$, contradicting the fact that v is nonzero. Therefore v_{n+1} is strictly positive. Now let $g = -\tilde{v}/v_{n+1}$, and by letting $y' = f(x')$, we get

$$f(x') \geq f(x) + \langle g, x' - x \rangle, \qquad \forall x' \in \mathcal{X},$$

showing that $g \in \partial f(x)$. □

A proof of the second part of Lemma (1.2) can be found in [Rockafellar, 1970, Theorem 25.1].

Exercise 1.6. Consider the function $f(x) = 1 - \sqrt{1 - x^2}$ for $x \in [-1, 1]$. Prove that $\partial f(-1) = \partial f(1) = \varnothing$. □

We shall also frequently use the following definitions, especially when analyzing the convergence rate of algorithms.

Definition 1.5. Let $f : \mathbb{R}^n \to \mathbb{R}$, and let $\mathcal{X} \subseteq \mathbb{R}^n$.

1. f is said to be G-Lipschitz continuous on $\mathcal{X}$, if

   $$|f(x) - f(y)| \leq G\|x - y\|$$

   for any $x, y \in \mathcal{X}$.
2. f is said to be L-smooth on $\mathcal{X}$, if f is differentiable on $\mathbb{R}^n$, and

   $$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

   for any $x, y \in \mathcal{X}$.

Some useful properties of Lipschitz continuous and smooth functions are summarized as follows.

Proposition 1.2 (Corollary of [Rockafellar, 1970, Theorem 24.7])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, and let $\mathcal{X} \subseteq \mathbb{R}^n$ be compact. Then there exists $G > 0$ such that*

1. *$f(x)$ is G-Lipschitz continuous on $\mathcal{X}$.*

2. $\|g\| \le G$ *for any* $g \in \partial f(x)$ *and any* $x \in \mathcal{X}$.

Proposition 1.3. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be differentiable. Then any one of the following conditions with* x, y *ranging over* $\mathbb{R}^n$ *is necessary and sufficient for* f *to be convex and* L*-smooth on* $\mathbb{R}^n$*:*

$$0 \le f(y) - (f(x) + \langle \nabla f(x), y - x\rangle) \le \frac{L}{2}\|y - x\|^2, \tag{1.6}$$

$$\|\nabla f(y) - \nabla f(x)\|^2 \le 2L(f(y) - (f(x) + \langle \nabla f(x), y - x\rangle)), \tag{1.7}$$

$$\|\nabla f(y) - \nabla f(x)\|^2 \le L\langle \nabla f(y) - \nabla f(x), y - x\rangle, \tag{1.8}$$

$$0 \le \langle \nabla f(y) - \nabla f(x), y - x\rangle \le L\|y - x\|^2. \tag{1.9}$$

Proof. The situation when $L = 0$ is trivial. We now assume $L > 0$.

- f is convex and smooth $\Rightarrow$ (1.6): Suppose f is convex and L-smooth on $\mathbb{R}^n$. The first inequality (1.6) follows from Proposition 1.1. To show the second inequality, by the Newton-Leibnitz Theorem, we get

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \langle \nabla f(x + t(y - x)), y - x\rangle \, dt \\ &= \langle \nabla f(x), y - x\rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x\rangle \, dt \\ &\le \langle \nabla f(x), y - x\rangle + \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\|\|y - x\| \, dt \\ &\le \langle \nabla f(x), y - x\rangle + \int_0^1 Lt\|y - x\|^2 \, dt \\ &= \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \end{aligned}$$

  which proves (1.6).

- (1.6) $\Rightarrow$ (1.7): Suppose (1.6) holds. We fix $x \in \mathbb{R}^n$ and construct $\phi(y) = f(y) - \langle \nabla f(x), y - x\rangle$. We then have, for any $y \in \mathbb{R}^n$,

$$\phi(y) = f(y) - \langle \nabla f(x), y - x\rangle \ge f(x) = \phi(x),$$

  showing that $\phi(y)$ achieves its global minimum over $\mathbb{R}^n$ at x. Consequently

$$\begin{aligned} \phi(x) &\le \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) = f\left(y - \frac{1}{L}\nabla\phi(y)\right) - \left\langle \nabla f(x), y - \frac{1}{L}\nabla\phi(y) - x\right\rangle \\ &\le f(y) + \left\langle \nabla f(y), -\frac{1}{L}\nabla\phi(y)\right\rangle + \frac{1}{2L}\|\nabla\phi(y)\|^2 - \left\langle \nabla f(x), y - \frac{1}{L}\nabla\phi(y) - x\right\rangle, \end{aligned}$$

  and by plugging in $\nabla\phi(y) = \nabla f(y) - \nabla f(x)$, we can get

$$\phi(x) \le f(y) - \langle \nabla f(x), y - x\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2,$$

  which is just (1.7).

- (1.7) $\Rightarrow$ (1.8): This follows by adding two copies of (1.7) with x and y interchanged.
- (1.8) $\Rightarrow$ f is convex and smooth: Suppose (1.8) holds. Since the norm is always nonnegative, we get $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0$, which, by Proposition 1.1, implies that f is convex on $\mathbb{R}^n$. Then we apply Cauchy-Schwarz inequality on the right-hand side of (1.8) and get

$$\|\nabla f(y) - \nabla f(x)\|^2 \leq L\|\nabla f(y) - \nabla f(x)\|\|y - x\|, \qquad \forall x, y \in \mathbb{R}^n,$$

  which implies $\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$ for all $x, y \in \mathbb{R}^n$. Therefore f is L-smooth on $\mathbb{R}^n$.
- (1.6) $\Rightarrow$ (1.9): This follows by adding two copies of (1.6) with x and y interchanged.
- (1.9) $\Rightarrow$ (1.6): Suppose (1.9) holds. By Proposition 1.1, the first inequality in (1.9) implies the first inequality in (1.6). Then, by Newton-Leibniz Theorem,

$$\begin{aligned} f(y) - (f(x) + \langle \nabla f(x), y - x \rangle) &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \, dt \\ &\leq \int_0^1 Lt\|y - x\|^2 \, dt = \frac{L}{2}\|y - x\|^2, \end{aligned}$$

  which proves the second inequality in (1.6).

The proof is now complete. □

Remark 1.1. Suppose $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set. If $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth on $\mathcal{X}$, then it's not hard to see from the above proof that (1.6) and (1.9) hold for any $x, y \in \mathcal{X}$. However, the converse is in general not true, i.e., (1.6) or (1.9) for any $x, y \in \mathcal{X}$ does not imply that f is L-smooth on $\mathcal{X}$ (though convexity can be implied). A counterexample is given by

$$f(x_1, x_2) = x_1 x_2, \qquad \forall (x_1, x_2) \in \mathbb{R}^2, \qquad \mathcal{X} = \{(x, 0) : x \in \mathbb{R}\}.$$

We then have

$$\nabla f(x_1, x_2) = (x_2, x_1),$$

and

$$f(y_1, 0) - f(x_1, 0) - \langle \nabla f(x_1, 0), (y_1, 0) - (x_1, 0) \rangle = 0,$$

showing that (1.6) holds for $L = 0$ for any $x, y \in \mathcal{X}$. On the other hand, f is obviously not 0-smooth on $\mathcal{X}$.

(Something to think about if you are really interested: What if $\mathcal{X}$ is convex and *open*?)

When the function f is further twice continuously differentiable, we can check the convexity and smoothness of f via the Hessian matrix of f.

Proposition 1.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable, and let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set with a nonempty interior. Then f is convex and L-smooth on $\mathcal{X}$ if and only if*

$$0 \preceq H_f(x) \preceq LI, \qquad \forall x \in \mathcal{X},$$

where $H_f(x)$ denotes the Hessian matrix of f at x.

Proof. • Necessity: Supose f is convex and L-smooth on $\mathcal{X}$. Let $x \in \operatorname{int}\mathcal{X}$ and $h \in \mathbb{R}^n$ be arbitrary. Denote $R_h(t) = \nabla f(x+th) - \nabla f(x) - t\,H_f(x)\,h$, and we have $\|R_h(t)\| = o(t)$ as $t \to 0$. Consequently, by (1.9),

$$0 \le \langle \nabla f(x+th) - \nabla f(x), th\rangle = t^2 h^\mathsf{T} H_f(x)\,h + t\langle h, R_h(t)\rangle \le Lt^2\|h\|^2.$$

We then divide the above inequality by t^2 and get

$$0 \le h^\mathsf{T} H_f(x)\,h + \frac{\langle h, R_h(t)\rangle}{t} \le L\|h\|^2.$$

By letting $t \to 0$ and noting that $\|R_h(t)\| = o(t)$, we get

$$0 \le h^\mathsf{T} H_f(x)\,h \le L\|h\|^2,$$

which holds for all $h \in \mathbb{R}^n$. Therefore by letting h be an eigenvector associated with eigenvalue $\lambda_{\min}(H_f(x))$ and $\lambda_{\max}(H_f(x))$ respectively, we get $\lambda_{\min}(H_f(x)) \ge 0$ and $\lambda_{\max}(H_f(x)) \le L$, which is equivalent to $0 \preceq H_f(x) \preceq LI$.

Now if $\mathcal{X}\setminus\operatorname{int}\mathcal{X}$ is nonempty, for any $x \in \mathcal{X}\setminus\operatorname{int}\mathcal{X}$, we can always find a sequence $(x_n)_{n\in\mathbb{N}}$ in $\operatorname{int}\mathcal{X}$ such that $x_n \to x$: Let $x_0 \in \operatorname{int}\mathcal{X}$ be arbitrary, and let

$$x_n = \frac{1}{n+1}x_0 + \left(1 - \frac{1}{n+1}\right)x.$$

Then obviously $x_n \in \mathcal{X}$ and $x_n \to x$. To show that $x_n \in \operatorname{int}\mathcal{X}$, let $\epsilon > 0$ be such that $x_0 + \epsilon v \in \mathcal{X}$ for any $v \in \mathbb{R}^n$ with $\|v\| < 1$. Then

$$x_n + \frac{1}{n+1}\epsilon v = \frac{1}{n+1}(x_0 + \epsilon v) + \left(1 - \frac{1}{n+1}\right)x \in \mathcal{X}$$

for any $\|v\| < 1$, showing that the open ball centered at x_n with radius $\epsilon/(n+1)$ is a subset of $\mathcal{X}$. Therefore $x_n \in \operatorname{int}\mathcal{X}$. Then $0 \preceq H_f(x) \preceq LI$ follows by the continuity of $H_f(x)$ over $x \in \mathbb{R}^n$.

- Sufficiency: Suppose $0 \preceq H_f(x) \preceq LI$ for all $x \in \mathcal{X}$. Then for any $x, y \in \mathcal{X}$, by the Newton-Leibniz rule, we have

$$\nabla f(y) - \nabla f(x) = \int_0^1 H_f(x + t(y-x))\,(y-x)\,dt,$$

and by taking the norm, we get

$$\begin{aligned}
\|\nabla f(y) - \nabla f(x)\| &= \left\| \int_0^1 H_f(x + t(y-x))\,(y-x)\,dt \right\| \\
&\le \int_0^1 \|H_f(x + t(y-x))\,(y-x)\|\,dt \\
&\le \int_0^1 \|H_f(x + t(y-x))\|_2\,\|(y-x)\|\,dt.
\end{aligned}$$

Since $0 \preceq H_f(x) \preceq LI$ for all $x \in \mathcal{X}$, we have $\|H_f(x + t(y - x))\|_2 \leq L$ for all $t \in [0, 1]$, which then implies

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|,$$

showing that f is L-smooth on $\mathcal{X}$. To show that f is convex on $\mathcal{X}$, we still start from $\nabla f(y) - \nabla f(x) = \int_0^1 H_f(x + t(y - x))\,(y - x)\,dt$ but this time we take the inner product:

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle = \int_0^1 (y - x)^\mathsf{T} H_f(x + t(y - x))\,(y - x)\,dt \geq 0,$$

where we used the fact that $H_f(x + t(y - x))$ is positive semidefinite for all $t \in [0, 1]$. We can now employ Proposition 1.1 to complete the proof. □

Exercise 1.7. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and μ-strongly convex on $\mathbb{R}^n$. Prove that

1. $\mu \leq L$, with equality only if $f(x) = \frac{L}{2}\|x\|^2 + q^\mathsf{T} x + r$, $\forall x \in \mathbb{R}^n$ for some $q \in \mathbb{R}^n$ and $r \in \mathbb{R}$.
2. The function $g(x) = f(x) - \frac{1}{2}\mu\|x\|^2$ is $(L - \mu)$-smooth on $\mathbb{R}^n$.

   (Hint: Show that the condition (1.9) holds for g.)
3. If f is further twice continuously differentiable, then

   $$\mu \leq \lambda_{\min}(H_f(x)) \leq \lambda_{\max}(H_f(x)) \leq L$$

   for all $x \in \mathbb{R}^n$, where $H_f(x)$ denotes the Hessian matrix of f at x.

   Conversely, show that, if f is twice continuously differentiable and $\mu \leq \lambda_{\min}(H_f(x)) \leq \lambda_{\max}(H_f(x)) \leq L$ for all $x \in \mathbb{R}^n$, then f is μ-strongly convex and L-smooth. □

1.3 Gradient Descent and Its Convergence Analysis

Consider the following unconstrained optimization problem:

$$\min_{x \in \mathbb{R}^d} \quad f(x)$$

where f is differentiable. The gradient descent (GD) method for solving this optimization problem is given by the following iteration:

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

There are different ways of choosing the step size η_t, and the following are some classic choices:

1. Constant step size $\eta_t = \eta$. This choice is convenient for convergence rate analysis. For example, it can be shown that when f is convex and L-smooth, then as long as $\eta \leq 1/L$, GD has a convergence rate guarantee $f(x_t) - f(x^*) = O(1/t)$ (see Theorem 1.3).
2. The Wolfe conditions: The Wolfe conditions consist of two conditions:

(a) The Armijo condition: The step size η_t should satisfy

$$f(x_t - \eta_t \nabla f(x_t)) \leq f(x_t) - c_1 \eta_t \|\nabla f(x_t)\|^2$$

for some constant $c_1 \in (0, 1)$. The Armijo condition guarantees that there will be sufficient decrease in the objective value.

(b) The curvature condition: The step size η_t should satisfy

$$\langle \nabla f(x_t - \eta_t \nabla f(x_t)), \nabla f(x_t) \rangle \leq c_2 \|\nabla f(x_t)\|^2$$

for some constant $c_2 \in (c_1, 1)$. The curvature condition rules out unacceptably short steps.

We shall not discuss more about the Wolfe conditions in this course. Interested readers may read [Nocedal and Wright, 2006, Section 3.1].

Exercise 1.8. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth. Show that, as long as $\eta_t \leq 2/L$, we have $f(x_{t+1}) \leq f(x_t)$, i.e., $f(x_t)$ is non-increasing in t. □

Exercise 1.9. Construct a "counterexample" in which gradient descent does not converge to the optimal solution. The counterexample should satisfy the following requirements:

1. The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and convex.
2. f attains its global minimum at some $x^* \in \mathbb{R}^n$.
3. $f(x_t)$, generated by gradient descent, is strictly decreasing.
4. $\sum_{t=0}^{\infty} \eta_t = +\infty$.
5. $f(x_t)$ does not converge to $f(x^*)$.

You may justify your construction of the example numerically. □

We now provide analysis of the convergence rate of GD for convex problems.

Theorem 1.3. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and convex, and the set of global minima*

$$\mathcal{X}^* = \left\{ x^* \in \mathbb{R}^d : f(x^*) = \inf_x f(x) \right\}$$

is nonempty. Then by choosing the step sizes to satisfy $\eta_t = \eta \in (0, 1/L]$, we have

$$f(x_t) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\eta t}$$

for all $t \geq 1$, where $x^* \in \mathcal{X}^*$ is arbitrary. Moreover, x_t converges to some point in $\mathcal{X}^*$.

Proof. Let $x^* \in \mathcal{X}^*$ be arbitrary, and denote $g_t = \nabla f(x_t)$. We have

$$\begin{aligned}\|x_t - x^*\|^2 &= \|x_{t+1} - x^* + \eta_t g_t\|^2 \\ &= \|x_{t+1} - x^*\|^2 + 2\eta_t \langle g_t, x_{t+1} - x^* \rangle + \eta_t^2 \|g_t\|^2 \\ &= \|x_{t+1} - x^*\|^2 - 2\eta_t \langle g_t, x^* - x_{t+1} \rangle + \|x_{t+1} - x_t\|^2,\end{aligned}$$

which leads to

$$\begin{aligned}&\frac{1}{2}\left(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right) \\ &= \eta_t \langle g_t, x^* - x_{t+1} \rangle - \frac{1}{2}\|x_{t+1} - x_t\|^2 \\ &= \eta_t \langle g_t, x^* - x_t \rangle + \eta_t \langle g_t, x_t - x_{t+1} \rangle - \frac{1}{2}\|x_{t+1} - x_t\|^2.\end{aligned} \tag{1.10}$$

Since f is convex and $g_t = \nabla f(x_t)$, we have

$$f(x_t) + \langle g_t, x^* - x_t \rangle \le f(x^*).$$

Therefore

$$\begin{aligned}&\frac{1}{2}\left(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right) \\ &\le \eta_t (f(x^*) - f(x_t)) + \eta_t \langle g_t, x_t - x_{t+1} \rangle - \frac{1}{2}\|x_{t+1} - x_t\|^2.\end{aligned} \tag{1.11}$$

Then, since f is L-smooth, we have

$$f(x_{t+1}) \le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2.$$

By using the above inequality to upper bound $-f(x_t)$ in (1.11), we get

$$\begin{aligned}&\frac{1}{2}\left(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right) \\ &\le \eta_t f(x^*) + \eta_t \left(-f(x_{t+1}) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2\right) \\ &\quad + \eta_t \langle g_t, x_t - x_{t+1} \rangle - \frac{1}{2}\|x_{t+1} - x_t\|^2 \\ &= \eta_t (f(x^*) - f(x_{t+1})) - \frac{1}{2}\left(1 - \eta_t L\right)\|x_{t+1} - x_t\|^2 \\ &\le \eta_t (f(x^*) - f(x_{t+1})),\end{aligned}$$

where in the last step we used $\eta_t L \le 1$. In other words,

$$\eta (f(x_{t+1}) - f(x^*)) \le \frac{1}{2}(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2). \tag{1.12}$$

By taking the telescoping sum, we get

$$\eta \sum_{\tau=1}^{t} (f(x_\tau) - f(x^*)) \le \frac{1}{2}(\|x_0 - x^*\|^2 - \|x_t - x^*\|^2) \le \frac{1}{2}\|x_0 - x^*\|^2.$$

Finally, we apply the result of Exercise 1.8 to obtain

$$f(x_t) - f(x^*) \le \frac{1}{t}\sum_{\tau=1}^{t} (f(\tau) - f(\tilde{(x)})) \le \frac{\|x_0 - x^*\|^2}{2\eta t},$$

which completes the convergence rate analysis.

To show that x_t converges to some point in $\mathcal{X}^*$, we note that (1.12) together with $f(x_{t+1}) \geq f(x^*)$ implies $\|x_{t+1}-x^*\|^2 \leq \|x_t-x^*\|^2$, showing that the sequence $(x_t)_{t\in\mathbb{N}}$ is bounded. Therefore there exist some $\tilde{x} \in \mathbb{R}^d$ and a subsequence $(x_{t_k})_{k\in\mathbb{N}}$ such that $x_{t_k} \to \tilde{x}$ as $k \to \infty$. By the L-smoothness of f, we have

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &= f(x_t) - \eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(x_t)\|^2, \end{aligned}$$

and by taking the limit $t \to \infty$ and noting $1 - \eta L/2 > 0$, we see that $\|\nabla f(x_t)\|^2 \to 0$ as $t \to \infty$. It then follows from the continuity of ∇f that

$$\nabla f(\tilde{x}) = \lim_{k\to\infty} \nabla f(x_{t_k}) = 0.$$

Therefore $\tilde{x} \in \mathcal{X}^*$, and we have $\|x_{t+1} - \tilde{x}\|^2 \leq \|x_t - \tilde{x}\|^2$ for all t. Now let $\epsilon > 0$ be arbitrary. Since $x_{t_k} \to \tilde{x}$ as $k \to \infty$, there exists $K \in \mathbb{N}$ such that $\|x_{t_K} - \tilde{x}\|^2 \leq \epsilon$, and consequently,

$$\|x_t - \tilde{x}\|^2 \leq \|x_{t_K} - \tilde{x}\|^2 \leq \epsilon$$

for all $t \geq t_K$. By the arbitrariness of $\epsilon > 0$, we see that $x_t \to \tilde{x} \in \mathcal{X}^*$ as $t \to \infty$. □

The following theorem shows that, the convergence rate of GD can be improved if f is strongly convex.

Theorem 1.4. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth. Then by choosing the step size to satisfy $\eta_t = \eta \in (0, 2/L)$, we have*

$$\|x_{t+1} - x^*\| \leq \max\{|1 - \eta\mu|, |\eta L - 1|\} \cdot \|x_t - x^*\|,$$

where x^* is the (unique) global minimizer of f over $\mathbb{R}^d$. Particularly, when $\eta = 2/(\mu + L)$, $\max\{|1 - \eta\mu|, |\eta L - 1|\}$ achieves the minimum value, and

$$\|x_{t+1} - x^*\| \leq \frac{L - \mu}{L + \mu}\|x_t - x^*\|.$$

Proof. Without loss of generality we assume that $\mu < L$. We shall prove a stronger statement:[1] For any $x, y \in \mathbb{R}^d$, we have

$$\|y^+ - x^+\| \leq \max\{|1 - \eta\mu|, |\eta L - 1|\} \cdot \|y - x\|,$$

where

$$x^+ = x - \eta\nabla f(x), \qquad y^+ = y - \eta\nabla f(y).$$

[1] This statement is equivalent to saying that the mapping $x \mapsto x - \eta\nabla f(x)$ is a contraction mapping over $\mathbb{R}^d$ with contraction coefficient $\max\{|1 - \eta\mu|, |\eta L - 1|\}$.

We have

$$\begin{aligned}\|y^+ - x^+\|^2 &= \|y - x - \eta(\nabla f(y) - \nabla f(x))\|^2 \\ &= \|y-x\|^2 - 2\eta\langle \nabla f(y) - \nabla f(x), y - x\rangle + \eta^2\|\nabla f(y) - \nabla f(x)\|^2.\end{aligned}$$

Fix x temporarily and let $\phi(z) = f(z) - \frac{1}{2}\mu\|z-x\|^2$. As Exercise 1.7 shows, ϕ is $(L-\mu)$-smooth. By applying (1.8) to ϕ, we get

$$\|\nabla f(y) - \nabla f(x) - \mu(y-x)\|^2 \le (L-\mu)\langle \nabla f(y) - \nabla f(x) - \mu(y-x), y - x\rangle,$$

which implies

$$\langle \nabla f(y) - \nabla f(x), y - x\rangle \ge \frac{\mu L}{\mu + L}\|y-x\|^2 + \frac{1}{\mu+L}\|\nabla f(y) - \nabla f(x)\|^2. \tag{1.13}$$

We now consider two cases:

1. $0 < \eta \le 2/(\mu + L)$. In this case, we let $L' = 2/\eta - \mu$, and it can be seen that $L' \ge L$. Therefore f is also L'-smooth, and (1.13) holds with L replaced by L'. Consequently,

$$\begin{aligned}\|y^+ - x^+\|^2 &\le \|y-x\|^2 - 2\eta\left(\frac{\mu L'}{\mu + L'}\|y-x\|^2 + \frac{1}{\mu+L'}\|\nabla f(y) - \nabla f(x)\|^2\right) \\ &\quad + \eta^2\|\nabla f(y) - \nabla f(x)\|^2 \\ &= \left(1 - \frac{2\eta\mu L'}{\mu + L'}\right)\|y-x\|^2 - \eta\left(\frac{2}{\mu+L'} - \eta\right)\|\nabla f(y) - \nabla f(x)\|^2 \\ &= (1-\eta\mu)^2\|y-x\|^2.\end{aligned}$$

   Note that in this case we have $1 - \eta\mu \ge |\eta L - 1|$.

2. $2/(\mu + L) < \eta < 2/L$. In this case, we let $\mu' = 2/\eta - L$, and it can be seen that $0 < \mu' < \mu$. Therefore f is also μ'-strongly convex, and (1.13) holds with μ replaced by μ'. Consequently, we have

$$\begin{aligned}\|y^+ - x^+\|^2 &\le \|y-x\|^2 - 2\eta\left(\frac{\mu' L}{\mu' + L}\|y-x\|^2 + \frac{1}{\mu'+L}\|\nabla f(y) - \nabla f(x)\|^2\right) \\ &\quad + \eta^2\|\nabla f(y) - \nabla f(x)\|^2 \\ &= \left(1 - \frac{2\eta\mu' L}{\mu' + L}\right)\|y-x\|^2 - \eta\left(\frac{2}{\mu'+L} - \eta\right)\|\nabla f(y) - \nabla f(x)\|^2 \\ &= (\eta L - 1)^2\|y-x\|^2.\end{aligned}$$

   Note that in this case we have $\eta L - 1 > |1 - \eta\mu|$.

Combining these two cases, we get the desired results. □

Exercise 1.10. Show that when f is μ-strongly convex and L-smooth, under the condition

$\eta_t = \eta \in (0, 2/L)$, the convergence rate of GD in terms of $f(x_t) - f(x^*)$ is given by

$$f(x_t) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2} \cdot \alpha^t,$$

where

$$\alpha = \max\{(1 - \eta\mu)^2, (\eta L - 1)^2\}. \qquad \square$$

We now provide some discussions:

1. The two theorems presented above analyze the convergence behavior of GD in terms of the *convergence rate*. The convergence rate is a quantitative characterization of how fast the algorithm's output approaches optimality, or how fast the sub-optimality of the algorithm's output decreases as the number of iterations increases. Specifically, in Theorem 1.3 and in Exercise 1.10, we use $f(x_t) - f(x^*)$ to quantify the sub-optimality of the output of the algorithm at step t, and the convergence rates describes how fast $f(x_t) - f(x^*)$ decreases as t increases.

   Apart from $f(x_t) - f(x^*)$, there are other metrics for quantifying the sub-optimality of the algorithm's output. Which metric is appropriate for theoretical analysis depends on the optimization algorithm itself and the properties of the problem to be solved.

   The convergence rate is one of the central notions of performance guarantees in theoretical studies of optimization algorithms.

2. The convergence rate is closely related with the notion of *iteration complexity*, which counts the number of iterations needed for achieving an arbitrary degree of optimality. For example, from Theorem 1.3, we can derive that, for an arbitrary $\epsilon > 0$, if the number of iterations t satisfies

   $$t \geq \frac{\|x_0 - x^*\|^2}{2\eta} \cdot \frac{1}{\epsilon} = O\left(\frac{1}{\epsilon}\right),$$

   then

   $$f(x_t) - f(x^*) \leq \epsilon.$$

   In other words, the number of iterations needed to achieve $f(x_t) - f(x^*) \leq \epsilon$ is on the order of $O(1/\epsilon)$, which gives the iteration complexity of GD under the conditions of Theorem 1.3.

   The iteration complexity is another critical notion for evaluating the efficiency of the algorithm. In general, the iteration complexity can be derived from the convergence rate. The iteration complexity can be more useful than the convergence rate if we are further interested in the scalability of the algorithm. For example, suppose the convergence rates of two algorithms are given by

   $$f(x_t) - f(x^*) \leq C \cdot \frac{d}{t} \quad \text{and} \quad f(x_t) - f(x^*) \leq C \cdot \sqrt{\frac{d}{t}}.$$

   We see that the right-hand sides have different dependences on the problem dimension d.

However, their iteration complexities are then

$$O\left(\frac{d}{\epsilon}\right) \quad \text{and} \quad O\left(\frac{d}{\epsilon^2}\right),$$

which indicates that the two algorithms in fact have similar (theoretical) scalability as the problem dimension d increases.

1.4 Stochastic Gradient Descent

The stochastic gradient descent (SGD) can be viewed as a variant of GD, which solves the unconstrained optimization problem $\min_{x\in\mathbb{R}^d} f(x)$ by the following iterations:

$$x_{t+1} = x_t - \eta_t\, G(x_t;\xi_t),$$

where $\xi_1, \xi_2, \ldots$ is a sequence of independent random variables, and $G(x;\xi_t)$ is a stochastic gradient satisfying

$$\mathbb{E}[G(x;\xi_t)] = \nabla f(x)$$

for all $x \in \mathbb{R}^d$. In other words, we replace the true gradient in GD by a random estimator which only equals the true gradient in expectation. Naturally, the stochastic gradient descent can be applied in situations where obtaining the true gradient is not possible or is expensive, but a stochastic gradient can be obtained relatively easily. A typical scenario is supervised learning with a large data set, which can be formulated as follows:

$$\min_{\theta\in\mathbb{R}^d} \quad R(\theta) = \frac{1}{D}\sum_{i=1}^{D} \ell(x_i, y_i; \theta).$$

Here $\{(x_i, y_i) : i = 1, \ldots, D\}$ is a labeled dataset with D being very large, θ represents the parameterized model, and $\ell(x, y; \theta)$ is the loss function that evaluates how the model θ fits the single data (x, y). The goal is to find the best model that minimizes the empirical risk $R(\theta)$. Since D is very large, it may be costly to evaluate the full gradient $\nabla R(\theta) = \frac{1}{D}\sum_{i=1}^{D} \nabla_\theta \ell(x_i, y_i; \theta)$. However, constructing a stochastic gradient is not hard: At each time step t, we choose a batch size B_t and randomly sample a small subset $\mathcal{B}_t \subseteq \{1, \ldots, D\}$ of size B_t from the uniform distribution. We then construct

$$G(\theta; \mathcal{B}_t) = \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t} \nabla_\theta \ell(x_i, y_i; \theta).$$

It's not hard to check that $\mathbb{E}[G(\theta; \mathcal{B}_t)] = \nabla R(\theta)$, and as long as the batch size B_t is small enough, evaluating $G(\theta; \mathcal{B}_t)$ is relatively easy. The SGD method is then given by

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta_t\, G(\theta_t; \mathcal{B}_t) \\ &= \theta_t - \eta_t \cdot \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t} \nabla_\theta \ell(x_i, y_i; \theta).\end{aligned}$$

In order to establish convergence results for SGD, we need to impose assumptions on the variance (or the second moment) of the stochastic gradient. The following is a commonly-used assumption for establishing the convergence rate of SGD.

Assumption 1.1. There exists $\sigma > 0$ such that

$$\mathbb{E}\big[\|G(x;\xi_t) - \nabla f(x)\|^2\big] \le \sigma^2$$

for all $x \in \mathbb{R}^d$ and all $t \in \mathbb{N}$.

We now derive the convergence rate of SGD.

Theorem 1.5. *Suppose f is convex and L-smooth, and suppose $G(x;\xi_t)$ is an unbiased estimator of $\nabla f(x)$ for each t and satisfies Assumption 1.1. Furthermore, suppose we choose the step sizes to satisfy $\eta_t \in (0, 1/(2L)]$. Then the iterates of the stochastic gradient descent method satisfies*

$$\mathbb{E}[f(\bar{x}_t) - f(x^*)] \le \frac{\frac{1}{2}\mathbb{E}\big[\|x_0 - x^*\|^2\big] + \sigma^2 \sum_{\tau=0}^{t-1} \eta_\tau^2}{\sum_{\tau=0}^{t-1} \eta_\tau},$$

where

$$\bar{x}_t = \frac{\sum_{\tau=1}^{t} \eta_{\tau-1} x_\tau}{\sum_{\tau=1}^{t} \eta_{\tau-1}}.$$

Proof. Denote $g_t = G(x_t;\xi_t)$. The identity (1.10) will still apply here:

$$\begin{aligned} &\frac{1}{2}\left(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right) \\ &= \eta_t\langle g_t, x^* - x_t\rangle + \eta_t\langle g_t, x_t - x_{t+1}\rangle - \frac{1}{2\eta_t}\|x_{t+1} - x_t\|^2. \end{aligned} \tag{1.10}$$

Notice that, since $\mathbb{E}[g_t|x_t] = \nabla f(x_t)$, we get

$$\mathbb{E}[f(x_t) + \langle g_t, x^* - x_t\rangle|x_t] = f(x_t) + \langle \nabla f(x_t), x^* - x_t\rangle \le f(x^*),$$

and consequently,

$$\begin{aligned} &\frac{1}{2}\mathbb{E}\big[\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\big|x_t\big] \\ &\le \eta_t(f(x^*) - f(x_t)) + \eta_t\mathbb{E}[\langle g_t, x_t - x_{t+1}\rangle|x_t] - \frac{1}{2}\mathbb{E}\big[\|x_{t+1} - x_t\|^2|x_t\big] \\ &= \eta_t(f(x^*) - f(x_t)) + \eta_t\mathbb{E}[\langle \nabla f(x_t), x_t - x_{t+1}\rangle|x_t] \\ &\quad + \eta_t\mathbb{E}[\langle g_t - \nabla f(x_t), x_t - x_{t+1}\rangle|x_t] - \frac{1}{2}\mathbb{E}\big[\|x_{t+1} - x_t\|^2|x_t\big] \end{aligned}$$

Next, by the L-smoothness of f, we have

$$f(x_{t+1}) \le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2,$$

which leads to

$$\begin{aligned} &\frac{1}{2}\mathbb{E}\big[\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\big|x_t\big] \\ &\le \eta_t f(x^*) + \eta_t\mathbb{E}\left[-f(x_{t+1}) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2\big|x_t\right] \end{aligned}$$

$$
\begin{aligned}
&+ \eta_t \mathbb{E}[\langle \nabla f(x_t), x_t - x_{t+1} \rangle | x_t] + \eta_t \mathbb{E}[\langle g_t - \nabla f(x_t), x_t - x_{t+1} \rangle | x_t] - \frac{1}{2} \mathbb{E}\big[\|x_{t+1} - x_t\|^2 | x_t\big] \\
&= \mathbb{E}\left[\eta_t (f(x^*) - f(x_{t+1})) - \frac{1}{2}(1 - \eta_t L) \|x_{t+1} - x_t\|^2 \middle| x_t \right] \\
&\quad + \eta_t \mathbb{E}[\langle g_t - \nabla f(x_t), x_t - x_{t+1} \rangle | x_t] \,.
\end{aligned}
$$

Moreover, by the inequality $2ab \le a^2/\theta + \theta b^2$, we see that[2]

$$
\begin{aligned}
\mathbb{E}[\langle g_t - \nabla f(x_t), x_t - x_{t+1} \rangle | x_t] &\le \frac{1}{2} \mathbb{E}\left[2\eta_t \|g_t - \nabla f(x_t)\|^2 + \frac{1}{2\eta_t} \|x_{t+1} - x_t\|^2 \middle| x_t \right] \\
&\le \eta_t \sigma^2 + \frac{1}{4\eta_t} \mathbb{E}\big[\|x_{t+1} - x_t\|^2 | x_t\big] \,.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
&\frac{1}{2} \mathbb{E}\big[\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2 \big| x_t\big] \\
&\le \mathbb{E}\left[\eta_t (f(x^*) - f(x_{t+1})) - \frac{1}{2} \left(\frac{1}{2} - \eta_t L \right) \|x_{t+1} - x_t\|^2 \middle| x_t \right] + \eta_t^2 \sigma^2 \\
&\le \mathbb{E}[\eta_t (f(x^*) - f(x_{t+1})) | x_t] + \eta_t^2 \sigma^2,
\end{aligned}
$$

where we used $\eta_t L \le 1/2$ for all t. In other words,

$$
\mathbb{E}[\, \eta_t (f(x_{t+1}) - f(x^*)) |\, x_t] \le \frac{1}{2} \mathbb{E}\big[\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 \big| x_t\big] + \eta_t^2 \sigma^2.
$$

By taking the total expectation and the telescoping sum, we get

$$
\mathbb{E}\left[\frac{\sum_{\tau=1}^t \eta_{\tau-1} (f(x_\tau) - f(x^*))}{\sum_{\tau=1}^t \eta_{\tau-1}} \right] \le \frac{\frac{1}{2} \mathbb{E}\big[\|x_0 - x^*\|^2\big] + \sigma^2 \sum_{\tau=0}^{t-1} \eta_\tau^2}{\sum_{\tau=0}^{t-1} \eta_\tau},
$$

and since the convexity of f implies

$$
f(\bar{x}_t) \le \frac{\sum_{\tau=1}^t \eta_{\tau-1} f(x_\tau)}{\sum_{\tau=1}^t \eta_{\tau-1}}
$$

for $\bar{x}_t = \big(\sum_{\tau=1}^t \eta_{\tau-1} x_\tau\big) / \big(\sum_{\tau=1}^t \eta_{\tau-1}\big)$, we get

$$
\mathbb{E}[f(\bar{x}_t) - f(x^*)] \le \frac{\frac{1}{2} \mathbb{E}\big[\|x_0 - x^*\|^2\big] + \sigma^2 \sum_{\tau=0}^{t-1} \eta_\tau^2}{\sum_{\tau=0}^{t-1} \eta_\tau},
$$

which is the desired result. □

[2] In fact, this step can be further strengthened if we note that for unconstrained SGD,

$$
\begin{aligned}
\mathbb{E}[\langle g_t - \nabla f(x_t), x_t - x_{t+1} \rangle | x_t] &= \eta_t \, \mathbb{E}[\langle g_t - \nabla f(x_t), g_t \rangle | x_t] \\
&= \eta_t \mathbb{E}\big[\|g_t - \nabla f(x_t)\|^2 | x_t\big] + \eta_t \mathbb{E}[\langle g_t - \nabla f(x_t), \nabla f(x_t) \rangle | x_t] \\
&= \eta_t \mathbb{E}\big[\|g_t - \nabla f(x_t)\|^2 | x_t\big] \le \eta_t \sigma^2.
\end{aligned}
$$

By pursuing this approach, the final convergence rate will be the same but the condition on the step size η_t will be weaker.

Corollary 1.1. *Suppose the conditions of Theorem 1.5 are satisfied.*

1. *If we choose the step sizes to satisfy* $\eta_t = c/(2L\sqrt{t+1})$ *for some* $c \in (0, 1]$*, then*

$$\mathbb{E}[f(\bar{x}_t) - f(x^*)] \leq O\left(\frac{\ln t}{\sqrt{t}}\right).$$

2. *If we plan ahead the total number of iterations to be* T*, and set the step sizes to be* $\eta_t = \eta = c/(2L\sqrt{T+1})$ *for some* $c \in (0, 1]$*, then*

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq O\left(\frac{1}{\sqrt{T}}\right),$$

and the iteration complexity to achieve $\mathbb{E}[f(\bar{x}_T) - f(x^*)]$ *is given by* $O(1/\epsilon^2)$*.*

For strongly convex and smooth objective functions, we have the following results on the convergence rate of SGD.

Theorem 1.6. *Suppose* f *is* μ*-strongly convex and* L*-smooth, and suppose* $G(x; \xi_t)$ *is an unbiased estimator of* $\nabla f(x)$ *for each* t *and satisfies Assumption 1.1. Let the step sizes be* $\eta_t = \frac{2}{\alpha(t+t_0)}$ *where* $\alpha \in (0, \mu]$ *and* $t_0 \geq 1$ *is sufficiently large such that* $\eta_t \leq 1/(2L)$ *for all* $t \in \mathbb{N}$*. Then the iterates of SGD satisfies*

$$f(\bar{x}_t) - f(x^*) \leq O\left(\frac{1}{t}\right),$$

where

$$\bar{x}_t = \sum_{\tau=1}^{t} \frac{2(\tau + t_0 - 2)}{t(t + 2t_0 - 3)} x_\tau$$

We postpone the proof of Theorem 1.6 to Appendix 1.A.

1.5 Other Variants of Gradient Descent

Subgradient descent. The subgradient descent method solves the unconstrained optimization problem $\min_{x \in \mathbb{R}^d} f(x)$ in which $f : \mathbb{R}^d \to \mathbb{R}$ is convex but potentially non-smooth. The iterations of the subgradient descent method is given by

$$x_{t+1} = x_t - \eta_t g_t,$$

where g_t is an arbitrary element of the subdifferential $\partial f(x_t)$.

Theorem 1.7. *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be a convex and* G*-Lipschitz continuous function. Let* x_t *be generatd by the subgradient descent method with step size* $\eta_t \propto 1/\sqrt{t+1}$*. Then*

$$f(\bar{x}_t) - f(x^*) \leq O\left(\frac{\ln t}{\sqrt{t}}\right), \qquad \text{where} \quad \bar{x}_t = \frac{\sum_{\tau=0}^{t-1} \eta_\tau x_\tau}{\sum_{\tau=0}^{t-1} \eta_\tau}.$$

Proof (Proof sketch). We can still start from (1.10):

$$\begin{aligned} &\frac{1}{2}\left(\|x_{t+1}-x^*\|^2-\|x_t-x^*\|^2\right) \\ &=\eta_t\langle g_t, x^*-x_t\rangle+\eta_t\langle g_t, x_t-x_{t+1}\rangle-\frac{1}{2}\|x_{t+1}-x_t\|^2. \end{aligned} \tag{1.10}$$

Also note that $f(x_t)+\langle g_t, x^*-x_t\rangle \le f(x^*)$ by the definition of subgradients. Furthermore, we have $\langle g_t, x_t-x_{t+1}\rangle \le \frac{\eta_t}{2}\|g_t\|^2+\frac{1}{2\eta_t}\|x_t-x_{t+1}\|^2 \le \frac{\eta_t}{2}G^2+\frac{1}{2\eta_t}\|x_t-x_{t+1}\|^2$. Summarizing these bounds, it can be shown that

$$\eta_t(f(x_t)-f(x^*)) \le \frac{1}{2}\left(\|x_t-x^*\|^2-\|x_{t+1}-x^*\|^2\right)+\frac{\eta_t^2G^2}{2}.$$

We can now take the telescoping sum and use Jensen's inequality to complete the proof. □

Projected gradient descent. Projected gradient decent (PGD) can be applied to constrained optimization problems of the form

$$\min_{x\in\mathcal{X}} \quad f(x)$$

where f is convex and smooth, and $\mathcal{X}\subseteq\mathbb{R}^d$ is a closed convex subset for which the projection operator $\mathcal{P}_{\mathcal{X}}$ is easy to compute. The iterations of PGD is given by

$$x_{t+1}=\mathcal{P}_{\mathcal{X}}[x_t-\eta_t\nabla f(x_t)].$$

When the true gradient $\nabla f(x_t)$ is replaced by a stochastic gradient $G(x_t;\xi_t)$, we get the projected stochastic gradient descent (PSGD) method:

$$x_{t+1}=\mathcal{P}_{\mathcal{X}}[x_t-\eta_t\, G(x_t;\xi_t)].$$

When f is assumed to be convex and L-smooth, the analysis of PGD or PSGD starts from the observation that

$$x_{t+1}=\mathcal{P}_{\mathcal{X}}[x_t-\eta_t g_t] \quad\Longleftrightarrow\quad \langle x_t-\eta_t g_t-x_{t+1}, z-x_{t+1}\rangle\le 0, \quad \forall z\in\mathcal{X}$$

(see Exercise 1.2). By letting $z=x^*$ and noting that $\|x_t-x_{t+1}\|^2+\|x^*-x_{t+1}\|^2-\|x^*-x_t\|^2=2\langle x_t-x_{t+1}, x^*-x_{t+1}\rangle$, we can derive the following inequality version of (1.10):

$$\begin{aligned} &\frac{1}{2}\left(\|x_{t+1}-x^*\|^2-\|x_t-x^*\|^2\right) \\ &\le\eta_t\langle g_t, x^*-x_t\rangle+\eta_t\langle g_t, x_t-x_{t+1}\rangle-\frac{1}{2}\|x_{t+1}-x_t\|^2. \end{aligned} \tag{1.14}$$

The subsequent steps are almost identical to those of GD or SGD.

Exercise 1.11. Let $f:\mathbb{R}^n\to\mathbb{R}$ be convex and L-smooth over $\mathcal{X}$, and let x_t be generated by the projected gradient descent iterations with step sizes $\eta_t=\eta\in(0,2/L)$. Show that

$f(x_t)$ is non-increasing.

(Hint: By the L-smoothness of f, we have

$$f(x_{t+1}) \le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2.$$

Then show that

$$\langle \nabla f(x_t), x_{t+1} - x_t \rangle \le -\frac{1}{\eta}\|x_{t+1} - x_t\|^2$$

by using the result of Exercise 1.2.) □

Exercise 1.12. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be closed and convex. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, and is μ-strongly convex and L-smooth on $\mathcal{X}$. Let x_t be generated by the projected gradient descent iterations with a constant step size $\eta_t = \eta \in (0, 2/L)$.

1. Show that $\|x_{t+1} - x^*\|^2 \le \|x_t - x^* - \eta(\nabla f(x_t) - \nabla f(x^*))\|^2$.

   (Hint: First prove that $x^* = \mathcal{P}_{\mathcal{X}}[x^* - \eta \nabla f(x^*)]$ for the constrained minimizer x^*. Then apply the nonexpansiveness of the projection operator.)

2. Derive the convergence rate of the projected gradient descent iterations. Note that this time, since f is strongly convex only on a subset $\mathcal{X}$ of $\mathbb{R}^d$, it may not be valid to use the inequality (1.13). For simplicity, you may assume that f is twice continuously differentiable over $\mathbb{R}^n$. □

Further generalizations of PGD and PSGD for constrained/composite optimization problems include the mirror descent (MD) method [Beck and Teboulle, 2003, Bubeck, 2015], the proximal gradient descent method [Parikh and Boyd, 2014], etc.

Nesterov's accelerated gradient descent. Nesterov's accelerated gradient descent is a (class of) first-order method that achieves faster convergence than the vanilla (i.e., original) gradient descent method for smooth convex optimization.

Consider the unconstrained minimization problem $\min_{x \in \mathbb{R}^d} f(x)$. When f is convex and L-smooth, a commonly used version of Nesterov's accelerated GD is given by

$$\begin{aligned} x_{t+1} &= y_t - \frac{1}{L}\nabla f(y_t), \\ y_{t+1} &= x_{t+1} + \frac{\alpha_{t+1}(1 - \alpha_t)}{\alpha_t}(x_{t+1} - x_t), \end{aligned} \tag{1.15}$$

where the sequence $\alpha_t \in (0, 1)$ is generated by arbitrarily choosing $\alpha_0 \in (0, 1)$ and solving $\alpha_{t+1}^2 = (1 - \alpha_{t+1})\alpha_t^2$, and we initialize with $y_0 = x_0$. When f is further μ-strongly convex for

	Convex & Lipschitz	Convex & smooth	Strongly convex & smooth
(Projected) (Sub)GD	$O\left(\frac{\ln t}{\sqrt{t}}\right)$	$O\left(\frac{1}{t}\right)$	$O\left((1-c\kappa)^t\right)$
Stochastic (Sub)GD	$O\left(\frac{\ln t}{\sqrt{t}}\right)$	$O\left(\frac{\ln t}{\sqrt{t}}\right)$	$O\left(\frac{1}{t}\right)$
Nesterov's method	N/A	$O\left(\frac{1}{t^2}\right)$	$O((1-\sqrt{\kappa})^t)$

Table 1.1: Convergence rates for some typical gradient-descent-type methods. Here $\kappa = \mu/L$ for μ-strongly convex and L-smooth objective functions, and $c > 0$ is some numerical constant. The rates are upper bounds of the optimality gap $\mathbb{E}[f(\bar{x}_t) - f(x^*)]$ where $\bar{x}_t$ is certain weighted average of $x_0, x_1, \ldots, x_t$ (see relevant theorems in the notes). Some results have not been proved in this set of notes, which we leave as possible exercises for interested readers. You may also consult relevant literature such as [Bubeck, 2015], [Nesterov, 2018], etc.

some $\mu > 0$, a commonly used version of Nesterov's accelerated GD is given by

$$\begin{aligned} x_{t+1} &= y_t - \frac{1}{L}\nabla f(y_t), \\ y_{t+1} &= x_{t+1} + \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}(x_{t+1} - x_t), \end{aligned} \tag{1.16}$$

where $\kappa = \mu/L$, and we initialize with $y_0 = x_0$.

The convergence of Nesterov's accelerated GD is given by the following theorem, whose proof can be found in [Nesterov, 2018, Section 2.2.1] (see Theorem 2.2.1 and Lemma 2.2.4 therein).

Theorem 1.8. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth, and suppose x^* is a minimizer of $f(x)$ over $x \in \mathbb{R}^d$.*

1. *Let x_t be generated by (1.15). Then*

$$f(x_t) - f(x^*) \le O\left(\frac{1}{t^2}\right).$$

2. *If f is further μ-strongly convex for some $\mu > 0$, and let x_t be generated by (1.16). Then*

$$f(x_t) - f(x^*) \le \left(1-\sqrt{\kappa}\right)^t \left(f(x_0) - f(x^*) + \frac{\mu}{2}\|x_0 - x^*\|^2\right),$$

where $\kappa = \mu/L$.

We summarize the convergence rate results of gradient-descent-type methods in Table 1.1, which will be used as centralized benchmarks for comparison with distributed optimization methods.

Exercise 1.13. Let $L > \mu > 0$ and $r > 0$, and define

$$f(x) = \sum_{i=1}^{p} g(a_i^\mathsf{T} x - b_i) + \frac{\mu}{2}\|x\|^2$$

for $x \in \mathbb{R}^n$, where

$$g(x) = \begin{cases} \frac{1}{2}x^2 e^{-r/x}, & x > 0, \\ 0, & x \le 0, \end{cases}$$

$b_i \in \mathbb{R}$ for each i, and $a_1, \ldots, a_p \in \mathbb{R}^n$ are column vectors of a matrix $A \in \mathbb{R}^{n\times p}$ satisfying $\|A\|_2 = \sqrt{L-\mu}$.

1. Show that f is L-smooth and μ-strongly convex.

   (Hint: Show that f is twice differentiable and $\mu I \preceq H_f(x) \preceq LI$, where $H_f(x)$ denotes the Hessian matrix of f at x.)

2. Numerical experiment: Fix $n = 100$, $p = 50$, $r = 10^{-6}$ and $L = 10^4$. Generate each b_i from the normal distribution $\mathcal{N}(0,1)$, and generate $A \in \mathbb{R}^{n\times p}$ by $A = \sqrt{L-\mu}\tilde{A}/\|\tilde{A}\|_2$ where each entry of $\tilde{A}$ is sampled from $\mathcal{N}(0,1)$. Set $x_0 = 0$, and run the vanilla gradient descent method

   $$x_{t+1} = x_t - \eta\nabla f(x_t), \qquad \eta = \frac{2}{\mu + L}$$

   and Nesterov's accelerated gradient descent method (1.16) to solve the problem

   $$\min_{x\in\mathbb{R}^n} \quad f(x)$$

   for $\mu = 100$, $\mu = 10$ and $\mu = 1$. Plot the values of $\log_{10}(f(x_t) - f(x^*))$ versus the iteration number t for the three cases. □

1.6 Basic Graph Theory

Undirected Graph

An *undirected graph* is an ordered pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a set of *nodes* or *vertices*, and $\mathcal{E}$ is the set of *edges* whose elements are of the form $\{i, j\}$ with $i, j \in \mathcal{V}$ and $i \neq j$.[3] An element $\{i, j\} \in \mathcal{E}$ represents an undirected edge connecting node i and node j. In this course, we always assume that $\mathcal{V}$ is a finite set, and, without loss of generality, assume $\mathcal{V} = \{1, \ldots, n\}$ for some $n \in \mathbb{N}$.

Now let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph. The following are some basic definitions in graph theory:

[3]The requirement $i \neq j$ excludes self-loops.

- We say that a node j is a *neighbor* of node i if $\{i,j\} \in \mathcal{E}$. The set of neighbors of node i will be denoted by $\mathcal{N}_i$, i.e., $\mathcal{N}_i \coloneqq \{j \in \mathcal{V} : \{i,j\} \in \mathcal{E}\}$.
- The *degree* of node i, denoted by $\deg(i)$, is the total number of neighbors of node i.

  The *degree matrix* of $\mathcal{G}$, denoted by D, is the $n \times n$ diagonal matrix whose i'th diagonal element is equal to $\deg(i)$.
- A *path* connecting node i and node j is a finite sequence of nodes $v_1, \ldots, v_m \in \mathcal{V}$ such that $v_1 = i$, $v_m = j$ and $\{v_k, v_{k+1}\} \in \mathcal{E}$ for each $k = 1, \ldots, m-1$. The number $m-1$ is called the *length* of the path. For a path $p = (v_1, \ldots, v_m)$ in $\mathcal{G}$ and an edge $e \in \mathcal{E}$, we sometimes use the notation $e \in p$ to mean that $e = \{v_k, v_{k+1}\}$ for some $k \in \{1, \ldots, m-1\}$.

  Let $i, j \in \mathcal{V}$ with $i \neq j$ be arbitrary. The *distance* from i to j is the minimum value of the lengths of all paths connecting i and j.
- We say that the undirected graph $\mathcal{G}$ is *connected*, if for any $i, j \in \mathcal{V}$ with $i \neq j$ there exists a path connecting i and j.

  When $\mathcal{G}$ is connected, we define its *diameter* as the maximum value of the distances from any $i \in \mathcal{V}$ to any $j \in \mathcal{V}$ with $i \neq j$.
- The *adjacency matrix* A of the graph $\mathcal{G}$ is the $n \times n$ matrix defined by

$$A_{ij} = \begin{cases} 1, & \text{if } \{i,j\} \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases}$$

  Obviously A is a real symmetric matrix.

Exercise 1.14. Let $\mathcal{G} = (\{1, \ldots, n\}, \mathcal{E})$ be an undirected graph, and let A be its adjancency matrix.

1. Let $i, j \in \{1, \ldots, n\}$ be arbitrary. When will the (i,j)'th entry of A^2 be positive?
2. Show that there exists a path of length k from node i to node j if and only if the (i,j)'th entry of A^k is positive.
3. Suppose that $n \geq 2$. Show that $\mathcal{G}$ is connected if and only if all entries of $\sum_{k=0}^{n-1} A^k$ are positive. □

- The *Laplacian matrix* L of the graph $\mathcal{G}$ is defined by

$$L_{ij} = \begin{cases} \deg(i), & \text{if } i = j, \\ -1, & \text{if } i \neq j \text{ and } \{i,j\} \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases}$$

  Obviously L is a real symmetric matrix, and we also have $L = D - A$.

Lemma 1.3. *Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ denotes the n eigenvalues of L. The following statements hold:*

1. *L is positive semidefinite.*
2. *$\lambda_n = 0$ and $L\mathbf{1} = 0$ where $\mathbf{1}$ is the vector whose entries are all equal to 1.*
3. *$\lambda_{n-1} > \lambda_n$ if $\mathcal{G}$ is connected.*

Proof. 1. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be arbitrary. We then have

$$\begin{aligned} 0 \leq \sum_{\{i,j\} \in \mathcal{E}} (x_i - x_j)^2 &= \frac{1}{2} \sum_{i,j=1}^{n} A_{ij}(x_i - x_j)^2 \\ &= \frac{1}{2} \sum_{i,j=1}^{n} A_{ij} x_i^2 + \frac{1}{2} \sum_{i,j=1}^{n} A_{ij} x_j^2 - \sum_{i,j=1}^{n} A_{ij} x_i x_j \\ &= \sum_{i=1}^{n} x_i^2 \sum_{j=1}^{n} A_{ij} - \sum_{i,j=1}^{n} A_{ij} x_i x_j \\ &= \sum_{i=1}^{n} \deg(i) \cdot x_i^2 - \sum_{i,j=1}^{n} A_{ij} x_i x_j = x^{\mathsf{T}} L x, \end{aligned}$$

which shows that L is positive semidefinite.

2. We have

$$(L\mathbf{1})_i = \sum_{j=1}^{n} L_{ij} = \deg(i) - \sum_{j=1}^{n} A_{ij} = 0.$$

3. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be any vector such that $Lx = 0$. We then have $x^{\mathsf{T}} L x = 0$, which, by the calculation in the first part of the proof, implies

$$\sum_{\{i,j\} \in \mathcal{E}} (x_i - x_j)^2 = 0,$$

which further implies $x_i - x_j = 0$ whenever $\{i, j\} \in \mathcal{E}$. Now let $p, q \in \mathcal{V}$ be arbitrary. Since $\mathcal{G}$ is connected, there exists a path $p = v_1, v_2, \ldots, v_m = q$ such that $\{v_k, v_{k+1}\} \in \mathcal{E}$ for each $k = 1, \ldots, m-1$. Therefore we have

$$x_p - x_{v_1} = x_{v_1} - x_{v_2} = \cdots = x_{v_{n-1}} - x_q = 0,$$

which leads to $x_p - x_q = 0$. By the arbitrariness of $p, q \in \mathcal{V}$, we see that $x = c\mathbf{1}$ for some $c \in \mathbb{R}$. Therefore the eigenvalue $\lambda_n = 0$ has multiplicity 1, and consequently $\lambda_{n-1} > \lambda_n$. □

Exercise 1.15. Let $\mathcal{G} = (\{1, \ldots, n\}, \mathcal{E})$ be an undirected graph. Suppose $\deg(i) > 0$ for all $i \in \mathcal{V}$. We define the *normalized Laplacian matrix* by

$$\mathcal{L} \coloneqq D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2},$$

where $D^{-1/2}$ is the diagonal matrix whose i'th diagonal element is $1/\sqrt{\deg(i)}$. Show that

1. For any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we have

$$x^\mathsf{T} \mathcal{L} x = \sum_{\{i,j\} \in \mathcal{E}} \left(\frac{x_i}{\sqrt{\deg(i)}} - \frac{x_j}{\sqrt{\deg(j)}} \right)^2 .$$

2. The smallest eigenvalue of $\mathcal{L}$ is 0, and has multiplicity 1 if $\mathcal{G}$ is connected.

3. Show that the largest eigenvalue of $\mathcal{L}$, denoted by $\lambda_{\max}(\mathcal{L})$, is given by

$$\lambda_{\max}(\mathcal{L}) = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\sum_{\{i,j\} \in \mathcal{E}} (x_i - x_j)^2}{\sum_{i \in \mathcal{V}} x_i^2 \cdot \deg(i)}.$$

   Then show that $\lambda_{\max}(\mathcal{L}) \leq 2$. As a byproduct,[a] the largest eigenvalue of the (unnormalized) Laplacian matrix L is upper bounded by $2 \max_{i \in \mathcal{V}} \deg(i)$. □

[a] Though it is not really necessary to resort to the normalized Laplacian $\mathcal{L}$ to derive an upper bound of $\lambda_{\max}(L)$.

Directed Graph

A *directed graph* (or *digraph*) is an ordered pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a set of *nodes* or *vertices*, and $\mathcal{E}$ is the set of *edges* whose elements are ordered pairs (i, j) with $i, j \in \mathcal{V}$. An element $(i, j) \in \mathcal{E}$ represents a directed edge *from i to j*. We allow self-loops in digraphs, i.e., $(i, i) \in \mathcal{E}$ represents an edge from node i to itself. We shall always assume $\mathcal{V}$ is a finite set, and, without loss of generality, assume $\mathcal{V} = \{1, \ldots, n\}$ for some $n \in \mathbb{N}$.

Now let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph.

- We say that a node j is an *in-neighbor* of node i if $(j, i) \in \mathcal{E}$. A node k is an *out-neighbor* of node i if $(i, k) \in \mathcal{E}$.
- The *in-degree* and *out-degree* of a node i, denoted by $\deg_{\text{in}}(i)$ and $\deg_{\text{out}}(i)$, are the numbers of in-neighbors and out-neighbors of i, respectively.
- A *path* connecting node i and node j is a finite sequence of nodes $v_1, \ldots, v_m \in \mathcal{V}$ such that $v_1 = i$, $v_m = j$ and $(v_k, v_{k+1}) \in \mathcal{E}$ for each $k = 1, \ldots, m-1$. For a path $p = (v_1, \ldots, v_m)$ in $\mathcal{G}$ and an edge $e \in \mathcal{E}$, We sometimes use the notation $e \in p$ to mean that $e = (v_k, v_{k+1})$ for some $k \in \{1, \ldots, m-1\}$.
- The graph $\mathcal{G}$ is called *strongly connected* if for any pair of nodes $i, j \in \mathcal{V}$ there exists a path from i to j.
- The *adjacency matrix* A of the graph $\mathcal{G}$ is the $n \times n$ matrix defined by

$$A_{ij} = \begin{cases} 1, & \text{if } (i, j) \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases}$$

  It can be seen that the entries of $A\mathbf{1}$ are the out-degrees of each node in $\mathcal{G}$, while $A^\mathsf{T}\mathbf{1}$ are the in-degrees of each node in $\mathcal{G}$.

For digraphs, the adjancency matrix is in general not symmetric. When A is symmetric, we have $(i,j) \in \mathcal{E}$ if and only if $(j,i) \in \mathcal{E}$, and in this case we say that the digraph $\mathcal{G}$ is undirected. It's not hard to see that for an undirected digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, there exists a digraph $\mathcal{G}' = (\mathcal{V}, \mathcal{E}')$ such that $\{i,j\} \in \mathcal{E}$ if and only if $i \neq j$ and $(i,j) \in \mathcal{E}$, and the adjacency matrix A' of $\mathcal{G}'$ is equal to A expect possibly for diagonal entries.

1.7 Basic Setups of Distributed Optimization

Broadly speaking, the term *distributed optimization* may refer to any technique of solving optimization problems via a group of agents connected by a communication network. Each agent is able to carry out certain computing tasks on its own, and is also able to exchange information with its neighbors in the communication network whenever necessary. In this course, we shall only discuss *discrete-time* algorithms, i.e., the time domain is discretized, and the optimization algorithms are specified by iterations in discrete-time.

Among all the possible configurations of agents and communication networks for distributed optimization, there are two basic setups that are prevalent in the research area of distributed optimization:

1. The server-worker setup: In this setup, there is one particular agent, which we call the server, that is responsible for collecting data and distributing tasks to other agents. Other agents are called workers, and each worker only communicates with the server in the communication network. Consequently, the communication network has a star topology, where the server is at the center node and the workers are at the surrounding nodes.
2. The peer-to-peer setup: In this setup, each agent plays roughly the same role in the optimization procedure, and there is no central coordinator (or an external central coordinator can broadcast to the agents but receives information from the agents very infrequently). The topology of the communication network can be arbitrary (and even time-varying), as long as it is "sufficiently connected" so that any piece of information can be spread to all nodes in the network eventually.

In this course, we will mainly focus on the peer-to-peer setup, and will particularly study the so-called *consensus optimization* problems. At the end of the course, we will briefly introduce federated learning from the perspective of distributed optimization under the server-worker setup.

Notes on References

[Boyd and Vandenberghe, 2004] gives a friendly introduction to the general theory of convex analysis and optimization. Readers with sufficient mathematical maturity may also consult the classic book [Rockafellar, 1970] on convex analysis. [Nesterov, 2018] and [Bubeck, 2015] present convergence rate/complexity analysis of fundamental convex optimization algorithms. The test case in Exercise 1.13 is adapted from [Van Scoy et al., 2017]; this paper also proposes the *triple momentum* method that achieves faster convergence rate than the standard Nesterov's accel-

erated GD for strongly convex functions. The recent monograph [d'Aspremont et al., 2021] provides a quite thorough and modern introduction to acceleration techniques in convex optimization. The materials of basic graph theory are mostly adopted from [Bullo, 2022].

The convergence rate of GD has also been studied via the tools of performance estimation problems (PEP) [Drori and Teboulle, 2014], and it has been shown that for GD with a constant step size η applied to an L-smooth convex function f, we have

$$f(x_t) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{4\eta L t + 2},$$

and the upper bound is tight in the sense that for any $L > 0, t \in \mathbb{N}$ and $d \in \mathbb{N}$, we can find an L-smooth function $\varphi : \mathbb{R}^d \to \mathbb{R}$ and an initial point $x_0 \in \mathbb{R}^d$ such that $\varphi(x_t) - \varphi(x^*) = L\|x_0 - x^*\|^2/(4\eta L t + 2)$. In addition, a series of very recent papers [Grimmer et al., 2023, Altschuler and Parrilo, 2023a, Altschuler and Parrilo, 2023b] have found that the convergence of the vanilla GD algorithm can be accelerated by using non-constant step sizes. For example, [Altschuler and Parrilo, 2023b] showed that, with properly chosen non-constant step sizes, the vanilla GD for unconstrained smooth convex problems can achieve the $O(1/\epsilon^\kappa)$ complexity bound, where $\kappa = \ln 2 / \ln(1 + \sqrt{2}) \approx 0.7864$.

1.A Proof of Theorem 1.6

The proof provided here is mostly adapted from [Bubeck, 2015]. Denote $g_t = G(x_t; \xi_t)$. We still start from (1.10):

$$\begin{aligned} &\frac{1}{2}\left(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right) \\ &= \eta_t \langle g_t, x^* - x_t \rangle + \eta_t \langle g_t, x_t - x_{t+1} \rangle - \frac{1}{2\eta_t}\|x_{t+1} - x_t\|^2. \end{aligned} \tag{1.10}$$

Note that, since $\alpha \geq \mu$, $f(x) - \frac{\alpha}{2}\|x - x^*\|^2$ is convex, and so we have

$$\begin{aligned} &\mathbb{E}[f(x_t) + \langle g_t, x^* - x_t \rangle | x_t] \\ &= f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle \\ &= f(x_t) - \frac{\alpha}{2}\|x_t - x^*\|^2 + \langle \nabla f(x_t) - \alpha(x_t - x^*), x^* - x_t \rangle - \frac{\alpha}{2}\|x_t - x^*\|^2 \\ &\leq f(x^*) - \frac{\alpha}{2}\|x_t - x^*\|^2. \end{aligned}$$

Then, by the L-smoothness of f, we have

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2.$$

By plugging these inequalities into (1.10) and taking the expectation conditioned on x_t, we can get

$$\mathbb{E}\left[f(x_{t+1}) - f(x^*) + \frac{1}{2\eta_t}\|x_{t+1} - x^*\|^2 \,\middle|\, x_t \right]$$

$$
\begin{aligned}
&\leq \left(\frac{1}{2\eta_t} - \frac{\alpha}{2}\right)\|x_t - x^*\|^2 + \mathbb{E}[\,\langle g_t - \nabla f(x_t), x_t - x_{t+1}\rangle \,|\, x_t] + \left(\frac{L}{2} - \frac{1}{2\eta_t}\right)\mathbb{E}\big[\,\|x_{t+1} - x_t\|^2 \big|\, x_t\big] \\
&\leq \left(\frac{1}{2\eta_t} - \frac{\alpha}{2}\right)\|x_t - x^*\|^2 + \eta_t \mathbb{E}\big[\,\|g_t - \nabla f(x_t)\|^2 \big|\, x_t\big] + \left(\frac{L}{2} - \frac{1}{4\eta_t}\right)\mathbb{E}\big[\,\|x_{t+1} - x_t\|^2 \big|\, x_t\big] \\
&\leq \left(\frac{1}{2\eta_t} - \frac{\alpha}{2}\right)\|x_t - x^*\|^2 + \eta_t \sigma^2,
\end{aligned}
$$

where in the second step we used $\langle g_t - \nabla f(x_t), x_t - x_{t+1}\rangle \leq \eta_t \|g_t - \nabla f(x_t)\|^2 + \|x_t - x_{t+1}\|^2/(4\eta_t)$, and the last step follows from $\eta_t \leq 1/(2L)$. By multiplying the above inequality with $t + t_0 - 1$, plugging in $\eta_t = \frac{2}{\alpha(t+t_0)}$ and taking the total expectation, we get

$$
\begin{aligned}
&(t + t_0 - 1)\,\mathbb{E}[(f(x_{t+1}) - f(x^*))] \\
&\leq \frac{2\sigma^2}{\alpha} + \frac{\alpha}{4}\left((t + t_0 - 1)(t + t_0 - 2)\,\mathbb{E}\big[\|x_t - x^*\|^2\big] - (t + t_0)(t + t_0 - 1)\,\mathbb{E}\big[\|x_{t+1} - x^*\|^2\big]\right).
\end{aligned}
$$

We can now take the telescoping sum and get

$$
\frac{2}{t(t + 2t_0 - 3)} \sum_{\tau=1}^{t} \mathbb{E}[(\tau + t_0 - 2)(f(x_\tau) - f(x^*))] \leq \frac{4\sigma^2}{\alpha(t + 2t_0 - 3)} + \frac{\alpha(t_0 - 1)(t_0 - 2)\|x_0 - x^*\|^2}{2t(t + 2t_0 - 3)}.
$$

The proof will be completed by applying Jensen's inequality.

Bibliography

[Altschuler and Parrilo, 2023a] Altschuler, J. M. and Parrilo, P. A. (2023a). Acceleration by stepsize hedging I: Multi-step descent and the silver stepsize schedule. *arXiv preprint arXiv:2309.07879*.

[Altschuler and Parrilo, 2023b] Altschuler, J. M. and Parrilo, P. A. (2023b). Acceleration by stepsize hedging II: Silver stepsize schedule for smooth convex optimization. *arXiv preprint arXiv:2309.16530*.

[Beck and Teboulle, 2003] Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

[Bubeck, 2015] Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

[Bullo, 2022] Bullo, F. (2022). *Lectures on Network Systems*. Kindle Direct Publishing, 1.6 edition.

[d'Aspremont et al., 2021] d'Aspremont, A., Scieur, D., and Taylor, A. (2021). Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245.

[Drori and Teboulle, 2014] Drori, Y. and Teboulle, M. (2014). Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1-2):451–482.

[Grimmer et al., 2023] Grimmer, B., Shu, K., and Wang, A. L. (2023). Accelerated gradient descent via long steps. *arXiv preprint arXiv:2309.09961*.

[Horn and Johnson, 2013] Horn, R. A. and Johnson, C. R. (2013). *Matrix Analysis*. Cambridge University Press, 2nd edition.

[Lax, 2007] Lax, P. D. (2007). *Linear Algebra and Its Applications*. John Wiley & Sons, 2 edition.

[Nesterov, 2018] Nesterov, Y. (2018). *Lectures on Convex Optimization*. Springer, 2 edition.

[Nocedal and Wright, 2006] Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, 2 edition.

[Parikh and Boyd, 2014] Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239.

[Rockafellar, 1970] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

[Rudin, 1976] Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill, 3 edition.

[Van Scoy et al., 2017] Van Scoy, B., Freeman, R. A., and Lynch, K. M. (2017). The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54.

Chapter 2

Consensus Optimization: Basics

2.1 Formulation and Applications

We first introduce the basic formulation of consensus optimization.

Consider a group of N agents, which are numbered by $i = 1, 2, \ldots, N$. We shall always assume that the number of agents N is greater than or equal to 2. The group of agents are connected by a communication network, which allows them to exchange information during the optimization procedure. Each agent is associated with a local cost function $f_i : \mathbb{R}^d \to \mathbb{R}$, and the goal is to solve the following optimization problem:

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{N} \sum_{i=1}^{N} f_i(x), \tag{2.1}$$

or equivalently,

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

The word "consensus" comes from the requirement that each agent should arrive at the same decision variable. We shall call the function

$$f(x) \coloneqq \frac{1}{N} \sum_{i=1}^{N} f_i(x)$$

the *global objective function* of the consensus optimization problem.

In consensus optimization, it is assumed that each agent i can obtain information (e.g, function values, gradients) on its local cost function f_i directly, but information on the local cost functions of other agents can only be obtained via exchanging information in the network.[1] Fur-

[1] This restriction does not mean that the formulation of consensus optimization cannot be applied to problems where some agents also have direct access to local cost functions of some other agents.

thermore, the time needed to transmit information from one agent to any of its neighbors is not negligible.

We now present some applications that can be modeled as consensus optimization problems.

Distributed learning. Suppose there are N agents, each of which has collected a set of labeled data $\mathcal{D}_i$ for supervised learning. Let $\ell(\theta; x, y)$ denote the loss function that quantifies how the model parameterized by $\theta \in \mathbb{R}^d$ fits the single pair of data (x, y). Let

$$f_i(\theta) = \sum_{(x,y)\in\mathcal{D}_i} \ell(\theta; x, y).$$

The empirical risk minimization problem for learning a model is can then be formulated as

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

Assume that, due to privacy concerns, the agents are not willing to share their data to other agents or upload their data to a central server, and need to learn the model by localized coordination via a communication network. Then we see that the distributed learning problem fits the formulation of consensus optimization.

Distributed state estimation. Consider a moving target whose state follows the linear dynamical system

$$x_{t+1} = A_t x_t + w_t,$$

where $(w_t)_{t\geq 0}$ is a sequence of i.i.d. process noises that follow the Gaussian distribution $\mathcal{N}(0, W)$ for some positive definite covariance matrix W. A group of N drones are connected by a communication network, and each drone i is equipped with a sensor that observes the moving target according to

$$y_{i,t} = C_{i,t} x_t + v_{i,t}$$

whenever the drone is sufficiently close to the target. Here $(v_{i,t})_{t\geq 0}$ is a sequence of i.i.d. measurement noises following the Gaussian distribution $\mathcal{N}(0, V_i)$ for some positive definite V_i for each $i = 1, \ldots, N$.

Now suppose the target moves from $t = 0$ to $t = T$, and each drone i observes the moving target for $t \in \mathcal{T}_i \subseteq \{0, \ldots, T\}$, and collects its observed data $(y_{i,t} : t \in \mathcal{T}_i)$. Assuming that A_t, W and the distribution of the initial state $x_0 \sim \mathcal{N}(\bar{x}_0, P_0)$ are known to all drones, and that each drone i also knows its $C_{i,t}$ and V_i. The goal for the group of drones is to estimate the trajectory $(x_0, \ldots, x_T)$ of the moving target using their collected measurement data, by solving the following optimization problem:

$$\min_{\hat{x}_0,\ldots,\hat{x}_T} \quad \|\hat{x}_0 - \bar{x}_0\|^2_{P_0^{-1}} + \sum_{t=1}^{T} \|\hat{x}_t - A_t\hat{x}_{t-1}\|^2_{W^{-1}} + \sum_{i=1}^{N}\sum_{t\in\mathcal{T}_i} \|y_{i,t} - C_{i,t}\hat{x}_t\|^2_{V_i^{-1}}.$$

A solution to the above optimization problem is the maximum a posteriori (MAP) estimator of the target's state. By letting

$$f_i(\hat{x}_0, \ldots, \hat{x}_T) = \|\hat{x}_0 - \bar{x}_0\|^2_{P_0^{-1}} + \sum_{t=1}^{T} \|\hat{x}_t - A_t\hat{x}_{t-1}\|^2_{W^{-1}} + N \sum_{t \in \mathcal{T}_i} \|y_{i,t} - C_{i,t}\hat{x}_t\|^2_{V_i^{-1}},$$

we see that the distributed state estimation problem satisfies the formulation of consensus optimization.

Distributed energy resource coordination. Consider the situation where a group of N distributed generators need to determine their power generations to meet the total demand while also minimizing the total generation cost. Mathematically, this problem can be formulated as

$$\begin{aligned} \min_{p_1,\ldots,p_i \in \mathbb{R}} \quad & \sum_{i=1}^{N} C_i(p_i) \\ \text{s.t.} \quad & \sum_{i=1}^{N} p_i = D, \\ & p_{i,\min} \le p_i \le p_{i,\max}. \end{aligned}$$

Here each $C_i : [0, +\infty) \to \mathbb{R}$ represents the generation cost of the ith distributed generator, which we assume is convex and continuous; D is the total demand, and $\sum_i p_i = D$ is the power balance constraint; $p_{i,\min}$ and $p_{i,\max}$ are the lower and upper limits of the power generation of the i'th generator.

We shall assume that $\sum_i p_{i,\min} < D < \sum_i p_{i,\max}$; in this case Slater's condition holds for this problem. Then we can formulate the dual problem:

$$\max_{\lambda \in \mathbb{R}} \quad \sum_{i=1}^{N} \phi_i(\lambda), \tag{2.2}$$

where

$$\phi_i(\lambda) = \min_{p_i \in [p_{i,\min}, p_{i,\max}]} C_i(p_i) - \lambda(p_i - D_i), \tag{2.3}$$

where D_i, $i = 1, \ldots, N$ are any real numbers satisfying $\sum_i D_i = D$. It's not hard to see that (2.2) satisfies the formulation of consensus optimization after rescaling and flipping the sign of the global objective. By the theory of convex analysis, it can be shown that, for any fixed $\lambda \in \mathbb{R}$, as long as p_i solves the optimization problem (2.3), we have

$$p_i \in \partial(-\phi_i)(\lambda).$$

Distributed routing control. Suppose there are N agents connected by a communication network. Each agent i is responsible for sending an amount of a certain type of commodity from one place (the source) to another place (the destination). We use a directed graph $\mathcal{G}_{\mathrm{f}} = (\mathcal{V}_{\mathrm{f}}, \mathcal{E}_{\mathrm{f}})$ to model the network of traffic, so that the sources s_i and the destinations t_i are all elements

of $\mathcal{V}_{\rm f}$, and each route that can be used by agent i for sending its commodity is a path in the network $\mathcal{G}_{\rm f}$ that connects s_i and v_i. Note that paths used by different agents are allowed to share edges in the network, so that each edge may be used to carry different types and amounts of commodities. We let $\mathcal{P}_i$ denote the set of paths that can be used by agent i. For agent i, the total traffic incurred for sending its commodity is given by $Q_i > 0$, and each agent needs to allocate this amount of traffic among the paths in $\mathcal{P}_i$. We let $x_{i,p}$ denote the *proportion* of traffic in Q_i allocated to the path $p \in \mathcal{P}_i$. We let x denote the vector that concatenates all $x_{i,p}$ for $p \in \mathcal{P}_i$ and $i = 1, \ldots, N$.

After each agent i determines its associated $x_{i,p}$ for each $p \in \mathcal{P}_i$, each edge $e \in \mathcal{E}$ will naturally carry a certain amount of traffic given by

$$q_e(x) = \sum_{i=1}^{N} \sum_{p: p \in \mathcal{P}_i \text{ and } e \in p} x_{i,p} Q_i,$$

i.e., we sum over all traffic that will go through the edge e. The traffic $q_e(x)$ will incur *congestion cost* for *each unit of traffic* along edge e, and it is given by $c_e(q_e(x))$ for some function $c_e : [0, +\infty) \to \mathbb{R}$. The local cost for agent i is then given by

$$f_i(x) = \sum_{p \in \mathcal{P}_i} \left(x_{i,p} Q_i \cdot \sum_{e \in p} c_e(q_e(x)) \right).$$

We emphasize that each agent's local cost may be affected by other agents' allocations of traffic, which introduces *coupling* among agents' decisions. Our goal is to find the optimal allocation of traffic that minimizes the global cost:

$$\begin{aligned} \min_{x} \quad & \frac{1}{N} \sum_{i=1}^{N} f_i(x), \\ \text{s.t.} \quad & x \in \bigcap_{i=1}^{N} \mathcal{X}_i \end{aligned}$$

where

$$\mathcal{X}_i = \left\{ x : \sum\nolimits_{p \in \mathcal{P}_i} x_{i,p} = 1,\, x \geq 0 \right\}.$$

Now we assume that the mapping $x \mapsto c_e(q_e(x))$ is known to agent i as long as there exists $p \in \mathcal{P}_i$ such that $e \in p$. Then this distributed routing control problem almost fits the formulation given by (2.1) except that it imposes local constraints $x \in \mathcal{X}_i$ for each i.

Exercise 2.1. Consider the distributed routing control problem. Show that the global objective function $f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x)$ satisfies

$$f(x) = \frac{1}{N} \sum_{e \in \mathcal{E}_{\rm f}} q_e(x) \cdot c_e(q_e(x)) \qquad \square$$

Exercise 2.2. Suppose there is a light source at location $z \in \mathbb{R}^3$, and around this light source there is a group of N sensors connected by a communication network. Each sensor i is located at $x_i \in \mathbb{R}^3$, and its measurement data y_i is given by the model

$$y_i = \frac{1}{\|x_i - z\|^{\beta_i}} + \alpha_i + w_i.$$

Here $\alpha_i \geq 0$ and $\beta_i > 0$ are constants known to agent i; w_i is a random measurement noise following the Gaussian distribution $\mathcal{N}(0, \sigma_i^2)$, where $\sigma_i > 0$ is also known to agent i. We assume that noises for different sensors are independent. We also assume that sensor i knows its location x_i.

Now let each sensor i take one measurement and obtain y_i for each $i = 1, \ldots, N$. Formulate the problem of finding the maximum likelihood estimator (MLE) of the light source's location z as a consensus optimization problem, and specify each local objective function f_i. □

Exercise 2.3. Suppose there are 3 agents, whose local objective functions are respectively $h_1(x_1, x_2)$, $h_2(x_2, x_3, x_4)$ and $h_3(x_4, x_5)$. The goal is to solve the following optimization problem:

$$\min_{x_i,\, i=1,\ldots,5} \quad h_1(x_1, x_2) + h_2(x_2, x_3, x_4) + h_3(x_4, x_5)$$

1. By introducing the variable $x = (x_1, x_2, x_3, x_4, x_5)$, reformulate the above problem as a consensus optimization problem with decision variable x. You may assume that each agent has sufficient information about the mapping from x to its own local objective value.
2. Now suppose h_1, h_2 and h_3 are each (jointly) strongly convex. We reformulate the original problem as

$$\begin{aligned} \min_{\substack{x_i,\, i=1,\ldots,5 \\ y_2, y_4}} \quad & h_1(x_1, x_2) + h_2(y_2, x_3, x_4) + h_3(y_4, x_5) \\ \text{s.t.} \quad & x_2 = y_2, \\ & x_4 = y_4. \end{aligned}$$

   Find its dual problem and formulate the dual problem as a consensus optimization problem. □

2.2 Consensus Method for Distributed Averaging

In this section, we present the consensus method for distributed averaging.

Suppose there is a group of N agents connected by a communication network. For simplicity, the communication network is assumed to be static and bi-directional, and its topology is given

by the undirected graph $\mathcal{G} = (\{1, \dots, N\}, \mathcal{E})$ which is connected. Each agent is associated with a vector $x_i \in \mathbb{R}^d$, and the goal is to compute the average $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$. Each agent is only allowed to exchange information with its neighbors in the communication network.

A natural idea for the distributed averaging problem is as follows: Each agent iteratively takes some weighted average of all the vectors the agent can collect, i.e., the vector stored by itself and the vectors collected from its neighbors in the communication network. Mathematically, this procedure can be written as

$$x_i(t+1) = \sum_{j=1}^N W_{ij}\, x_j(t) \tag{2.4}$$

with the initialization $x_i(0) = x_i$. The weights W_{ij} should satisfy the following conditions:

1. The weights should be compatible with the topology of the communication network, i.e., $W_{ij} = 0$ whenever $i \neq j$ and $\{i, j\} \notin \mathcal{E}$.
2. $\sum_{j=1}^N W_{ij} = 1$ for all i.
3. $\sum_{i=1}^N W_{ij} = 1$ for all j.

The second condition ensures that, if the initial iterates have already achieved consensus, i.e., $x_i(0) = x_j(0)$ for all i, j, then $x_i(t)$ remains unchanged for all $t \geq 1$. The third condition ensures that, the mean of the vectors $x_i(t)$, $i = 1, \dots, N$ is preserved through the iterations, which can be seen from

$$\frac{1}{N}\sum_{i=1}^N x_i(t+1) = \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N W_{ij}x_j(t) = \frac{1}{N}\sum_{j=1}^N x_j(t)\sum_{i=1}^N W_{ij} = \frac{1}{N}\sum_{j=1}^N x_j(t).$$

Note that (2.4) can also be written as

$$x_i(t+1) = x_i(t) + \sum_{j\in\mathcal{N}_i} W_{ij}(x_j(t) - x_i(t)).$$

The iterations (2.4) give the *consensus method* for distributed averaging.

Remark 2.1. In order for the agents to implement (2.4) precisely, all agents must complete their required communication and computation for one step of (2.4) before they can proceed to the next iteration. Such methods are called *synchronous* methods. Some drawbacks of synchronized methods are as follows:

1. To run synchronous methods, it is usually required that all agents' local clocks are synchronized so that they can update at the same time. However, synchronization of local clocks can be hard for some communication networks.
2. Synchronous methods require each transmission of information to be perfect. This requirement is again hard to achieve for some communication networks.

On the other hand, synchronous methods are in general easier to analyze, and serve as bases for the design of most asynchronous algorithms.

We shall introduce the *weight matrix* $W \in \mathbb{R}^{N\times N}$ whose (i,j)'th entry is just W_{ij}.[2] The second and the third conditions are then equivalent to that $\mathbf{1}$ is an eigenvector of W and W^{T} with eigenvalue 1.

The three conditions provided previously are not sufficient for establishing convergence of the iterations (2.4) to the average $\bar{x}$. To study the convergence, we first introduce the notation

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

Then the iterations (2.4) can be equivalently written as

$$\mathsf{X}(t+1) = W\mathsf{X}(t). \tag{2.5}$$

We then note that

$$\bar{x}^{\mathsf{T}} = \frac{1}{N}\mathbf{1}^{\mathsf{T}}\mathsf{X}(0) = \frac{1}{N}\mathbf{1}^{\mathsf{T}}\mathsf{X}(t),$$

where we used the fact that the mean of $x_i(t)$ remains unchanged. Consequently, the deviation from the mean can be represented by

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

We now look at how the deviation from the mean evolves:

$$\begin{aligned} \mathsf{E}(t+1) &= \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right) W\mathsf{X}(t) \\ &= \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}W\right)\mathsf{X}(t) \\ &= \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)\mathsf{X}(t), \end{aligned}$$

while we also notice that

$$\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)\left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right) = W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}.$$

Therefore

$$\mathsf{E}(t+1) = \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)\mathsf{E}(t).$$

We now arrive at the following theorem:

[2] Other terminologies, such as *consensus matrix*, *mixing matrix*, *gossip matrix*, etc., have also been used in the literature.

Theorem 2.1. *Let $W \in \mathbb{R}^{N\times N}$ be a weight matrix satisfying $W\mathbf{1} = W^{\mathsf{T}}\mathbf{1} = 1$. If*

$$\rho\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right) < 1,$$

then the iterates of the consensus method (2.4) *satisfy*

$$\lim_{t\to\infty} \|x_i(t) - \bar{x}\| = 0.$$

Furthermore, if

$$\sigma := \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right\|_2 < 1,$$

then

$$\frac{1}{N}\sum_{i=1}^{N} \|x_i(t) - \bar{x}\|^2 \le \sigma^{2t} \cdot \frac{1}{N}\sum_{i=1}^{N} \|x_i(0) - \bar{x}\|^2 = O(\sigma^{2t})$$

As a corollary, we have the following result on the iteration complexity of the consensus method for distributed averaging.

Corollary 2.1. *Suppose the weight matrix W satisfies $W\mathbf{1} = W^{\mathsf{T}}\mathbf{1} = \mathbf{1}$ and*

$$\sigma = \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right\|_2 < 1.$$

Let $\epsilon > 0$ be arbitrary. Then the number of iterations T needed to achieve

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

satisfies

$$T = O\left(\frac{\ln(1/\epsilon)}{\ln(1/\sigma)}\right).$$

Exercise 2.4. Suppose $W \in \mathbb{R}^{N\times N}$ satisfies $Wu = u$ and $W^{\mathsf{T}}v = v$ for some $u, v \in \mathbb{R}^N\backslash\{0\}$ with $v^{\mathsf{T}}u \neq 0$. Show that

$$\left(W - \frac{uv^{\mathsf{T}}}{v^{\mathsf{T}}u}\right)^k = W^k - \frac{uv^{\mathsf{T}}}{v^{\mathsf{T}}u}.$$

Then show that

$$\lim_{k\to\infty} W^k = \frac{uv^{\mathsf{T}}}{v^{\mathsf{T}}u}$$

if and only if $\rho\big(W - \frac{uv^{\mathsf{T}}}{v^{\mathsf{T}}u}\big) < 1$. □

Exercise 2.5. Let $W \in \mathbb{R}^{N\times N}$ satisfy $W\mathbf{1} = W^\mathsf{T}\mathbf{1} = \mathbf{1}$ and $\rho\big(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\big) < 1$. We define the *asymptotic convergence factor* as

$$r_{\mathrm{asym}}(W) \coloneqq \sup_{x\in\mathbb{R}^N\setminus \mathrm{span}\{\mathbf{1}\}} \lim_{t\to\infty} \left(\frac{\big\|W^t x - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x\big\|}{\big\|x - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x\big\|}\right)^{1/t},$$

and the *per-step convergence factor* as

$$r_{\mathrm{step}}(W) \coloneqq \sup_{x\in\mathbb{R}^N\setminus \mathrm{span}\{\mathbf{1}\}} \frac{\big\|Wx - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x\big\|}{\big\|x - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x\big\|}.$$

Prove that

$$r_{\mathrm{asym}}(W) = \rho\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right) \qquad \text{and} \qquad r_{\mathrm{step}}(W) = \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2. \qquad \square$$

2.3 How to Construct the Weight Matrix

In this section, we present some approaches for constructing a weight matrix W that is compatible with the topology of the network and satisfies the conditions in Theorem 2.1.

Laplacian-Based Construction

We first provide a construction based on the Laplacian matrix.

Proposition 2.1. *Let $\mathcal{G} = (\{1,\dots,N\},\mathcal{E})$ be a connected undirected graph, and let L be the Laplacian matrix of $\mathcal{G}$. Let*

$$W = I - \alpha L,$$

where α is any real number satisfying $\alpha < \frac{2}{\lambda_{\max}(L)}$. Then

1. *W is compatible with the graph $\mathcal{G}$;*
2. *$W\mathbf{1} = W^\mathsf{T}\mathbf{1} = \mathbf{1}$;*
3. *$\big\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\big\|_2 < 1$.*

Proof. Let $\lambda_1 \geq \cdots \geq \lambda_{N-1} > \lambda_N = 0$ be the N eigenvalues of L. By the definition of the Laplacian matrix, it's evident that W is compatible with the graph $\mathcal{G}$. Then, $W\mathbf{1} = W^\mathsf{T}\mathbf{1} = \mathbf{1}$ follows from $L\mathbf{1} = 0$ and that W is real symmetric. Since $\mathcal{G}$ is connected, the eigenspace of L associated with the eigenvalue $\lambda_N = 0$ is spanned by $\{\mathbf{1}\}$, meaning that eigenspace of W associated with the eigenvalue 1 is spanned by $\{\mathbf{1}\}$. As a result, the eigenvalue decomposition of $W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}$ can be written as

$$W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} = \sum_{i=1}^{N-1}(1-\alpha\lambda_i)u_i u_i^\mathsf{T},$$

where $Wu_i = (1 - \alpha\lambda_i)u_i$ and $\{u_1, \dots, u_{N-1}\}$ forms an orthonormal basis of $(\operatorname{span}\{\mathbf{1}\})^\perp$. Thus, $\left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 < 1$ if and only if $1 - \alpha\lambda_i$ for all $i = 1, \dots, N-1$ have magnitudes strictly less than 1. The condition $\alpha < \frac{2}{\lambda_1}$ then implies

$$\max_{i=1,\dots,N-1} |1 - \alpha\lambda_i| = \max\{1 - \alpha\lambda_{N-1}, \alpha\lambda_1 - 1\} < 1,$$

which completes the proof. □

Remark 2.2. By Exercise 1.15, we have

$$\lambda_{\max}(L) \le 2 \max_{i=1,\dots,N} \deg(i),$$

which can be used to choose α that satisfies $\alpha < \frac{2}{\lambda_{\max}(L)}$ when computing the spectrum of L is difficult.

Exercise 2.6. Construct an example where the weight matrix W satisfies $W\mathbf{1} = W^\mathsf{T}\mathbf{1} = \mathbf{1}$ and $\left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 < 1$ but at least one of its entries is negative. The example can be constructed numerically.

(Hint: The diagonal elements of $I - \alpha L$ are given by $1 - \alpha \deg(i)$, which can be negative if $\alpha > 1/\deg(i)$ for some i.) □

Perron–Frobenius Theorem and the Metropolis Weights

We then discuss how to construct the weight matrix based on the Perron–Frobenius theorem.

Theorem 2.2 (Perron–Frobenius)**.** *Let $N \ge 2$, and let $P \in \mathbb{R}^{N\times N}$ have nonnegative entries. Let $\mathcal{G} = (\{1, \dots, N\}, \mathcal{E})$ be the directed graph whose adjacency matrix is given by*

$$A_{ij} = \begin{cases} 1, & \text{if } P_{ij} > 0, \\ 0, & \text{if } P_{ij} = 0. \end{cases}$$

Suppose the following conditions hold:

1. *$\mathcal{G}$ is strongly connected.*
2. *For each $i \in \{1, \dots, N\}$, the greatest common divisor of all lengths of paths in $\mathcal{G}$ that both start and end at node i is equal to 1.*

Then

1. *$\rho(P)$ is a simple eigenvalue of P and is strictly positive.*
2. *There is a unique $u \in \mathbb{R}^N$ such that $Pu = \rho(P)u$ and $\mathbf{1}^\mathsf{T}u = 1$; this vector has strictly positive entries.*
3. *All eigenvalues of P other than $\rho(P)$ have magnitudes strictly less than $\rho(P)$.*
4. *No eigenvectors of P except positive multiples of u have entries that are all nonnegative.*

As a corollary, we have the following result.

Proposition 2.2. *Let $\mathcal{G} = (\{1, \dots, N\}, \mathcal{E})$ be a connected undirected graph, and let $W \in \mathbb{R}^{N \times N}$ satisfy the following conditions:*

1. *W is a* doubly stochastic matrix*, i.e., $W_{ij} \geq 0$ for all i, j and $W\mathbf{1} = W^{\mathsf{T}}\mathbf{1} = \mathbf{1}$.*
2. *$W_{ii} > 0$ for all i.*
3. *For $i \neq j$, we have $W_{ij} > 0$ if and only if i and j are neighbors in $\mathcal{G}$.*

Then

$$\left\| W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}} \right\|_2 < 1.$$

Proof. Let $P = W^{\mathsf{T}}W$. It can be seen that P is positive semidefinite and $P\mathbf{1} = \mathbf{1}$, and all entries of P are nonnegative. Moreover,

$$P_{ii} = \sum_{k=1}^{N} W_{ki}W_{ki} \geq W_{ii}^2 > 0,$$

and for $i \neq j$ with $\{i, j\} \in \mathcal{E}$, we have

$$P_{ij} = \sum_{k=1}^{N} W_{ki}W_{kj} \geq W_{ii}W_{ij} > 0.$$

Therefore, if we let $A \in \mathbb{R}^{N \times N}$ be given by

$$A_{ij} = \begin{cases} 1, & \text{if } P_{ij} > 0, \\ 0, & \text{if } P_{ij} = 0, \end{cases}$$

and let $\mathcal{G}' = (\{1, \dots, N\}, \mathcal{E}')$ be the directed graph whose adjacency matrix is given by A, we see that $(i, j) \in \mathcal{E}'$ whenever $i = j$ or $\{i, j\} \in \mathcal{E}$. Since the (undirected) graph $\mathcal{G} = (\{1, \dots, N\}, \mathcal{E})$ is connected, we see that the (directed) graph $\mathcal{G}'$ is strongly connected. Furthermore, $(i, i) \in \mathcal{E}'$ for all $i = 1, \dots, N$ implies that the greatest common divisor of all the lengths of paths in $\mathcal{G}'$ that start and end at node i is equal to 1 for any $i = 1, \dots, N$. We can now apply the Perron–Frobenius theorem to conclude the following:

1. $\rho(P) = 1$.
2. 1 is a simple eigenvalue of P, with eigenvector $\mathbf{1}$ being an eigenvector.
3. All eigenvalues of P other than 1 are strictly less than 1.

Consequently, all eigenvalues of $P - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}$ lie in $[0, 1)$. Finaly, we note that

$$\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)^{\mathsf{T}}\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right) = P - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}},$$

and the proof will be completed by using the fact that the spectral norm of $W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}$ is equal to the square root of the maximum eigenvalue of $\big(W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\big)^{\mathsf{T}}\big(W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\big)$. □

Proposition 2.2 frees us from checking the eigenvalues of the weight matrix. We now provide one explicit example of weight matrices that can be used for the consensus method.

Example 2.1 (Metropolis weights)**.** Let $\mathcal{G} = (\{1, \dots, N\}, \mathcal{E})$ be a connected undirected graph, and let

$$W_{ij} = \begin{cases} \dfrac{1}{\max\{\deg(i), \deg(j)\} + \epsilon}, & \text{if } \{i,j\} \in \mathcal{E}, \\ 0, & \text{if } i \neq j \text{ and } \{i,j\} \notin \mathcal{E}, \\ 1 - \sum_{k \neq i} W_{ik}, & \text{if } i = j, \end{cases}$$

where ϵ is an arbitrary positive real number. It's not hard to check that W satisfies the conditions in Proposition 2.2, and therefore $\left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 < 1$. This weight matrix is called *Metropolis weight matrix*. □

Exercise 2.7. Let $N \geq 3$. Find the Metropolis weight matrix W and its corresponding $\sigma \coloneqq \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2$ for the following graphs. For the second and the third parts, also calculate the corresponding leading terms of the Taylor expansions of $\ln(1/\sigma)$ in N^{-1}:

1. A complete graph with N nodes.
2. A loop with N nodes, i.e., $\mathcal{G} = (\{1, \dots, N\}, \mathcal{E})$ with $\{i,j\} \in \mathcal{E}$ if and only if $|i - j| = 1$ or $i = 1, j = N$ or $i = N, j = 1$.

   (Hint: To compute $\left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2$, note that W is real symmetric and circulant, so that its eigenvalues can be found by the discrete Fourier transform. By definition, an $n \times n$ circulant matrix takes the form

   $$\begin{bmatrix} v & Tv & T^2v & \cdots T^{n-1}v \end{bmatrix}$$

   where $v \in \mathbb{R}^n$ is an arbitrary vector and $T : \mathbb{R}^n \to \mathbb{R}^n$ is the operator satisfying $(Tv)_1 = v_n$ and $(Tv)_i = v_{i-1}$ for $i = 2, 3, \dots, n$.)
3. (Optional) A 2D grid on a torus with N nodes where $N = K^2$ for some positive integer K. Specifically, the set of nodes is $\mathcal{V} = \{(i,j) : i,j \in \{1, \dots, K\}\}$, and $\{(i_1, j_1), (i_2, j_2)\} \in \mathcal{E}$ if and only if one of the following is satisfied:
   - $i_1 = i_2$ and $|j_1 - j_2| = 1$;
   - $j_1 = j_2$ and $|i_1 - i_2| = 1$;
   - $i_1 = i_2$, and either $j_1 = 1$, $j_2 = K$ or $j_1 = K$, $j_2 = 1$;
   - $j_1 = j_2$, and either $i_1 = 1$, $i_2 = K$ or $i_1 = K$, $i_2 = 1$. □

Finding Optimal Per-Step Convergence Factor

Let us suppose that the topology of the communication network $\mathcal{G} = (\{1, \dots, N\}, \mathcal{E})$ is known. We now study how to find the optimal W that minimizes the per-step convergence factor

$r_{\text{step}}(W) = \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2$ with the constraints that W is compatible with $\mathcal{G}$ and that $W\mathbf{1} = W^\mathsf{T}\mathbf{1} = \mathbf{1}$. In other words, we consider the following optimization problem to determine W:

$$\begin{aligned} \min_{W\in\mathbb{R}^{N\times N}} \quad & \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 \\ \text{s.t.} \quad & W_{ij} = 0,\ \forall i,j \text{ such that } i \neq j \text{ and } \{i,j\} \notin \mathcal{E}, \\ & W\mathbf{1} = \mathbf{1},\ \ W^\mathsf{T}\mathbf{1} = \mathbf{1}. \end{aligned} \tag{2.6}$$

This optimization problem is convex but nonsmooth in general. Next we will show how to transform this problem into a semidefinite program (SDP), which can then be solved by existing solvers.

We first note that, the convexity of the spectral norm implies

$$\begin{aligned} \left\|\frac{W+W^\mathsf{T}}{2} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 &= \left\|\frac{1}{2}\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right) + \frac{1}{2}\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)^\mathsf{T}\right\|_2 \\ &\leq \frac{1}{2}\left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 + \frac{1}{2}\left\|\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)^\mathsf{T}\right\|_2 = \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2. \end{aligned}$$

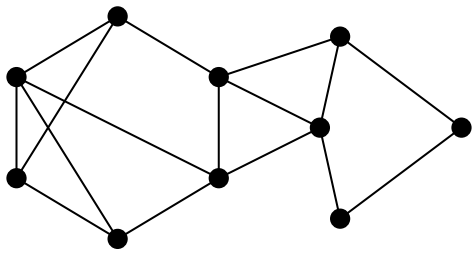
Since $\mathcal{G}$ is an undirected graph, $\frac{1}{2}\left(W+W^\mathsf{T}\right)$ will be compatible with $\mathcal{G}$ whenever W is compatible with $\mathcal{G}$. Moreover, $\frac{1}{2}\left(W+W^\mathsf{T}\right)\mathbf{1} = \mathbf{1}$ whenever $W\mathbf{1} = W^\mathsf{T}\mathbf{1} = \mathbf{1}$. As a result, the optimal value of (2.6) can always be achieved by a symmetric weight matrix W, which indicates that we may restrict W to be symmetric without losing generality.

Now, it is known that for any real symmetric matrix A, we have $\|A\|_2 \leq s$ if and only if $-sI \preceq A \preceq sI$. We can therefore formulate the following optimization problem for finding the optimal symmetric W:

$$\begin{aligned} \min_{W\in\mathbb{R}^{n\times n}, s\in\mathbb{R}} \quad & s \\ \text{s.t.} \quad & W_{ij} = 0,\ \forall i,j \text{ such that } i \neq j \text{ and } \{i,j\} \notin \mathcal{E}, \\ & W\mathbf{1} = \mathbf{1},\ \ W^\mathsf{T} = W, \\ & -sI \preceq W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} \preceq sI. \end{aligned} \tag{2.7}$$

This problem is an SDP, which can be numerically solved by existing SDP solvers or by interior-point methods tailored for the particular structures of the SDPs, when the graph $\mathcal{G}$ has small to medium scales. For very large networks, solving these SDPs can be prohibitive. We refer to [Xiao and Boyd, 2004] and relevant literature for more details and discussions.

Exercise 2.8. Consider the following graph:



Find $\left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2$ of the following weight matrices for the graph:

1. The Metropolis weight matrix with $\epsilon = 1$.
2. $W = I - \alpha L$, with $\alpha = 1/\max_i \deg(i)$.
3. $W = I - \alpha L$, where α is chosen to minimize $\left\|I - \alpha L - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2$.
4. The optimal solution to (2.7). □

Exercise 2.9. Let W be a symmetric and positive semidefinite weight matrix such that $W\mathbf{1} = \mathbf{1}$ and $\sigma \coloneqq \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 < 1$. We denote $A \coloneqq I - W$.

1. Show that A is positive semidefinite, $A\mathbf{1} = 0$, and 0 is a simple eigenvalue of A.
2. Consider the iteration

$$z(t+1) = Wz(t), \qquad \mathsf{x}(0) \in \mathbb{R}^N.$$

   Show that $(z(t))_{t\geq 0}$ coincides with the sequence generated by the gradient descent method for minimizing $\frac{1}{2}z^\mathsf{T}Az$. What is the corresponding step size?
3. Let $U_\perp \in \mathbb{R}^{N\times(N-1)}$ be a matrix satisfying $\mathbf{1}^\mathsf{T}U_\perp = 0$ and $U_\perp^\mathsf{T}U_\perp = I$. Consider the matrix

$$\tilde{A} \coloneqq U_\perp^\mathsf{T}AU_\perp.$$

   Show that the function $f(z) = \frac{1}{2}z^\mathsf{T}\tilde{A}z$ is $(1-\sigma)$-strongly convex and 1-smooth.

   Furthermore, let $\tilde{z}(t) \coloneqq U_\perp^\mathsf{T}z(t)$. Show that $\|\tilde{z}(t)\| = \|(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T})z(t)\|$, and that $(\tilde{z}(t))_{t\geq 0}$ coincides with the sequence generated by the gradient descent method for minimizing $\frac{1}{2}\tilde{z}^\mathsf{T}\tilde{A}\tilde{z}$. □

2.4 Extension to Directed Networks

We now consider the situation where the communication network is not bidirectional. As a consequence, we need to model the network by a digraph $\mathcal{G} = (\{1,\dots,N\},\mathcal{E})$. Specifically, $(i,j) \in \mathcal{E}$ for two distinct nodes i,j means that there is a (unidirectional) communication link via which node i is able to send information directly to node j. Without loss of generality, we assume that $\mathcal{G}$ does not have self-loops.

It's not hard to see that, as long as there exists a weight matrix $W \in \mathbb{R}^{N\times N}$ that is compatible with the digraph $\mathcal{G}$ (i.e., $W_{ij} = 0$ whenever $i \neq j$ and $(j,i) \notin \mathcal{E}$) and satisfies $W\mathbf{1} = W^{\mathsf{T}}\mathbf{1} = \mathbf{1}$ and $\rho\big(W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\big) < 1$, the previously presented theory can still be applied with minor modifications, and the iterations (2.4) will achieve convergence $x_i(t) \to \frac{1}{N}\sum_{i=1}^{N} x_i(0)$.

Optimizing Per-Step Convergence Factor for Digraphs. When the communication network is not bidirectional, we may still consider minimizing the per-step convergence factor $\|W - \mathbf{1}\mathbf{1}^{\mathsf{T}}/N\|_2$ provided that the digraph $\mathcal{G}$ is given. We use the following result: A matrix A satisfies $\|A\|_2 \leq s$ if and only if

$$\begin{bmatrix} sI & A \\ A^{\mathsf{T}} & sI \end{bmatrix} \succeq 0.$$

This result can be proved by using Schur complement. Then we can formulate the following optimization problem:

$$\begin{aligned} \min_{W\in\mathbb{R}^{N\times N}, s\in\mathbb{R}} \quad & s \\ \text{s.t.} \quad & W_{ij} = 0,\ \forall i,j \text{ such that } i \neq j \text{ and } (j,i) \notin \mathcal{E}, \\ & W\mathbf{1} = \mathbf{1},\ \ W^{\mathsf{T}}\mathbf{1} = \mathbf{1}, \\ & \begin{bmatrix} sI & W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}} \\ W^{\mathsf{T}} - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}} & sI \end{bmatrix} \succeq 0. \end{aligned} \tag{2.8}$$

It's not hard to see that the above optimization problem is an SDP. We refer to [Gharesifard and Cortés, 2010] for the existence of a feasible solution to (2.8).

However, the above procedure requires knowledge of the topology of the whole communication network, which may not be available in many practical scenarios. In general, when global information of the network's topology is not available to the agents, we need to design sophisticated coordination schemes for the agents to find a good set of weights for the consensus method (2.4) in a distributed fashion.

To address this issue, we introduce another approach to handle distributed averaging on digraphs. We start from the following lemma that considers relatively general weight matrices:

Lemma 2.1. *Let $W \in \mathbb{R}^{N\times N}$ be a weight matrix, and suppose 1 is a simple eigenvalue of W. Let $u, v \in \mathbb{C}^N$ satisfy $Wu = u$ and $W^{\mathsf{T}}v = v$. Then*

1. *$v^{\mathsf{T}}u \neq 0$.*
2. *Suppose further that all eigenvalues of W other than 1 have magnitudes strictly less than 1. Let $\mathsf{Y}(t) \in \mathbb{R}^{N\times d}$ be generated iteratively by $\mathsf{Y}(t+1) = W\mathsf{Y}(t)$. Then*

$$\lim_{t\to\infty} \mathsf{Y}(t) = \frac{uv^{\mathsf{T}}}{v^{\mathsf{T}}u}\mathsf{Y}(0).$$

Proof. 1. We consider the Jordan canonical form of the matrix W. Since 1 is a simple eigenvalue of W, we can write

$$S^{-1}WS = J = \begin{bmatrix} 1 & 0 \\ 0 & \Lambda \end{bmatrix},$$

where $S \in \mathbb{C}^{N\times N}$ is invertible, and $\Lambda \in \mathbb{C}^{(N-1)\times(N-1)}$ is an upper triangular matrix whose diagonals are all eigenvalues of W excluding 1. It's not hard to see from the equality $WS = SJ$ that the first column S is a right eigenvector of W with eigenvalue 1, and thus without loss of generality we may write

$$S = \begin{bmatrix} u & \tilde{S} \end{bmatrix}, \qquad \tilde{S} \in \mathbb{C}^{N\times(N-1)}.$$

Similarly, we can see from $S^{-1}W = JS^{-1}$ that the first row of S^{-1} is a left eigenvector of W with eigenvalue 1, and thus S^{-1} can be written in the form

$$S^{-1} = \begin{bmatrix} cv^{\mathsf{T}} \\ \check{S} \end{bmatrix}, \qquad \check{S} \in \mathbb{C}^{(N-1)\times N}.$$

where $c \in \mathbb{C}$ is a constant that can be determined as follows: By $S^{-1}S = I_N$, we have

$$\begin{bmatrix} cv^{\mathsf{T}} \\ \check{S} \end{bmatrix} \begin{bmatrix} u & \tilde{S} \end{bmatrix} = \begin{bmatrix} cv^{\mathsf{T}}u & cv^{\mathsf{T}}\tilde{S} \\ \check{S}u & \check{S}\tilde{S} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & I_{N-1} \end{bmatrix}.$$

Therefore we must have $cv^{\mathsf{T}}u = 1$, which implies that $v^{\mathsf{T}}u \neq 0$ and $c = (v^{\mathsf{T}}u)^{-1}$. As a useful byproduct, we can also obtain $\check{S}\tilde{S} = I_{N-1}$.

2. We continue our analysis from the first part. Notice that

$$W = SJS^{-1} = \begin{bmatrix} u & \tilde{S} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0\Lambda & \end{bmatrix} \begin{bmatrix} (v^{\mathsf{T}}u)^{-1}v^{\mathsf{T}} \\ \check{S} \end{bmatrix} = \frac{uv^{\mathsf{T}}}{v^{\mathsf{T}}u} + \tilde{S}\Lambda\check{S},$$

which leads to

$$\left(W - \frac{uv^{\mathsf{T}}}{v^{\mathsf{T}}u}\right)^k = \left(\tilde{S}\Lambda\check{S}\right)^k = \tilde{S}\Lambda^k\check{S},$$

where in the last step we used $\check{S}\tilde{S} = I_{N-1}$. Now, since all eigenvalues of W other than 1 have magnitudes strictly less than 1, all diagonal elements of Λ (and thus all eigenvalues of Λ) will also have magnitudes strictly less than 1, implying that $\Lambda^k \to 0$ as $k \to \infty$. We therefore get

$$\lim_{k\to\infty} \left(W - \frac{uv^{\mathsf{T}}}{v^{\mathsf{T}}u}\right)^k = 0,$$

and by the results of Exercise 2.4, we have

$$\lim_{k\to\infty} W^k = \frac{uv^{\mathsf{T}}}{v^{\mathsf{T}}u}.$$

Finally, we notice $\mathsf{Y}(t) = W^t\mathsf{Y}(0)$ and complete the proof. □

Exercise 2.10. Suppose $W \in \mathbb{R}^{N\times N}$ satisfies $Wu = u$ and $W^{\mathsf{T}}v = v$ for some $u, v \in$

$\mathbb{C}^N \backslash \{0\}$. Show that, if

$$v^\mathsf{T} u \neq 0 \quad \text{and} \quad \rho\left(W - \frac{uv^\mathsf{T}}{v^\mathsf{T} u}\right) < 1,$$

then 1 is a simple eigenvalue of W, and all other eigenvalues have magnitudes strictly less than 1. (This is the converse of Lemma 2.1.) □

We see that, if W satisfies the conditions stated in Lemma 2.1 and $W^\mathsf{T}\mathbf{1} = \mathbf{1}$, then for the sequence generated by the iterations

$$y_i(t+1) = \sum_{j=1}^N W_{ij} y_j(t), \qquad y_i(0) = x_i,$$

we have

$$\lim_{t\to\infty} y_i(t) = \frac{Nu_i}{\mathbf{1}^\mathsf{T} u} \cdot \bar{x} =: y_i(\infty),$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$ as usual. This equality means that agent i can recover $\bar{x}$ from $y_i(\infty)$ if it knows the quantity $Nu_i/\mathbf{1}^\mathsf{T} u$. But obtaining $Nu_i/\mathbf{1}^\mathsf{T} u$ is not a difficult task: We can introduce auxiliary variables $z_i(t) \in \mathbb{R}$ for each agent i, which are generated by the iterations

$$z_i(t+1) = \sum_{j=1}^N W_{ij} z_j(t), \qquad z_i(0) = 1.$$

Then

$$\lim_{t\to\infty} z_i(t) = \frac{Nu_i}{\mathbf{1}^\mathsf{T} u}.$$

We can now summarize the above derivations and propose the following iterations for distributed averaging over digraphs:

$$\begin{aligned} y_i(t+1) &= \sum_{j=1}^N W_{ij} y_j(t), \qquad y_i(0) = x_i, \\ z_i(t+1) &= \sum_{j=1}^N W_{ij} z_j(t), \qquad z_i(0) = 1, \\ x_i(t) &= \frac{y_i(t)}{z_i(t)}. \end{aligned} \tag{2.9}$$

Theorem 2.3. *Suppose $W \in \mathbb{R}^{N\times N}$ satisfies $W^\mathsf{T}\mathbf{1} = \mathbf{1}$, and 1 is a simple eigenvalue of W. Let $x(t)$ be generated by the iterations (2.9). Then*

$$\lim_{t\to\infty} x_i(t) = \frac{1}{N}\sum_{i=1}^N x_i$$

as long as all eigenvalues of W other than 1 have magnitudes strictly less than 1.

The final step is then to construct a weight matrix W that satisfies the conditions in Theorem 2.3. This is handled by the following proposition:

Proposition 2.3. *Suppose the digraph $\mathcal{G}$ is strongly connected and does not have self-loops. Let*

$$W_{ij} = \begin{cases} \dfrac{1}{1+\deg_{\text{out}}(j)}, & i = j \text{ or } (j,i) \in \mathcal{E}, \\ 0, & i \neq j \text{ and } (j,i) \notin \mathcal{E}. \end{cases} \tag{2.10}$$

Then,

1. *W is compatible with the digraph $\mathcal{G}$;*
2. *$W^{\mathsf{T}}\mathbf{1} = \mathbf{1}$;*
3. *1 is a simple eigenvalue of W;*
4. *All eigenvalues of W other than 1 have magnitudes strictly less than 1.*

Exercise 2.11. Prove Proposition 2.3. It should be clear which theorem of matrix analysis should be used. □

The iterations (2.9) with weights given by (2.10) are sometimes called *radio consensus*. Note that the construction of the weights in (2.10) only requires local structural information of the communication network.

2.5 Our First Distributed Optimization Algorithm

We return to the (unconstrained) consensus optimization problem

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

For simplicity, we assume that all agents are connected by a bidirectional communication network. We also assume that each f_i is sufficiently well behaved (e.g., convex and L-smooth).

If we ignore the restrictions imposed by the distributed setting, we may apply the gradient descent method

$$x(t+1) = x(t) - \eta \cdot \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x(t))$$

to obtain a good solution. In the distributed setting, the above iterations cannot be implemented directly. We let each agent i maintain a *local copy* of the decision variable, denoted by $x_i(t)$, with the requirement that all $x_i(t)$ are (approximately) equal to each other. We then have

$$x(t+1) = \frac{1}{N} \sum_{i=1}^{N} \left(x_i(t) - \eta \nabla f_i(x_i(t)) \right).$$

Note that the right-hand side is an average of local variables among all agents, which suggests we may apply the consensus method for distributed averaging to compute this quantity. Summarizing these ideas, we obtain our first distributed optimization algorithm given by Algorithm 1. Some remarks on this algorithm are as follows:

1. Lines 1 to 7 are intended for reaching a consensus on the first iterate $x_i(0)$. These steps can be removed if the initial value already satisfies $x_{i,0} = x_{j,0}$ for all i, j (e.g., if all agents initialize with $x_{i,0} = 0$).
2. Algorithm 1 operates on two time scales: Each iteration of the slow time scale carries out one gradient descent step, while the fast time scale iterations run the consensus method for distributed averaging. The t'th slow time scale iteration consists of K_{t+1} fast time scale iterations.
3. In Algorithm 1, we use coordinated step size η and numbers of inner iterations K_t for all agents. When designing distributed optimization algorithms, we usually start from a prototype in which each agent shares a coordinated set of algorithmic parameters, and then either assumes that there are prior distributed procedures that coordinate the algorithmic parameters for all agents, or try to generalize the algorithm with uncoordinated algorithmic parameters. For Algorithm 1, one naive approach to coordinate the step size is to run the following procedure:
   (a) Initialize $\eta_i(0) > 0$;
   (b) Run the following iterations
   $$\eta_i(t+1) = \min_{j \in \mathcal{N}_i \cup \{i\}} \eta_j(t)$$
   until t reaches the diameter of the graph.

   It's evident that the above procedure selects the minimum step size among $\eta_1(0), \ldots, \eta_N(0)$. The coordination of the numbers of inner iterations K_t can be implemented by distributed averaging with the help of the sufficient conditions in Theorem 2.4.

The rationale behind Algorithm 1 is in fact quite straightforward, given that we are now familiar with the consensus method for distributed averaging. A rigorous analysis of the convergence rate/complexity of Algorithm 1, however, requires some technical tricks that will be taught in later chapters. Here we only present the final results; the proofs are postponed to the appendices, which can be safely skipped.

Theorem 2.4. *Suppose that each f_i is L-smooth, that $f = \frac{1}{N}\sum_{i=1}^N f_i$ is convex with a global minimizer $x^* \in \mathbb{R}^d$, and that there exists $G > 0$ such that*

$$\|\nabla f_i(x) - \nabla f(x)\| \le G, \qquad \forall x \in \mathbb{R}^d, i \in \{1, \ldots, N\}.$$

Let $W \in \mathbb{R}^{N \times N}$ be a weight matrix satisfying $W\mathbf{1} = W^\mathsf{T}\mathbf{1} = \mathbf{1}$ and

$$\sigma \coloneqq \left\| W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} \right\|_2 < 1.$$

Algorithm 1: Our First Distributed Optimization Algorithm

Input: Weight matrix $W \in \mathbb{R}^{N\times N}$, step size $\eta > 0$, initial point $x_{i,0} \in \mathbb{R}^d$ for agent $i = 1, \dots, N$, non-decreasing sequence of positive integers $K_0, K_1, K_2, \dots$

Agent i sets $y_i(0;0) = x_{i,0}$
for $k = 0$ **to** $K_0 - 1$ **do**
  Agent i sends $y_i(0;k)$ to its neighbors
  Agent i updates
    $y_i(0;k+1) \leftarrow \sum_{j=1}^{N} W_{ij}\, y_j(0;k)$
end
Agent i sets $x_i(0) = y_i(0;K_0)$
for $t = 0, 1, 2, \dots$ **do**
  Agent i sets
    $y_i(t+1;0) = x_i(t) - \eta \cdot \nabla f_i(x_i(t))$
  for $k = 0$ **to** $K_{t+1} - 1$ **do**
    Agent i sends $y_i(t+1;k)$ to its neighbors
    Agent i updates
      $y_i(t+1;k+1) = \sum_{j=1}^{N} W_{ij}\, y_j(t+1;k)$
  end
  Agent i sets $x_i(t+1) = y_i(t+1;K_{t+1})$
end

Let $x_1, \dots, x_T$ be generated by Algorithm 1 with initialization $x_{i,0} = 0$ for all i, $\eta \in (0, 1/L]$ and

$$K_t \geq \left\lceil \frac{\ln[4((t+1)^2+3)]}{\ln(1/\sigma)} \right\rceil.$$

Then,

$$f\left(\frac{1}{t}\sum_{\tau=1}^{t} \bar{x}(\tau)\right) - f(x^*) \leq \frac{\left(\|x^*\| + 2\eta^2 LG\right)^2}{2\eta t}, \qquad \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|x_i(t) - \bar{x}(t)\|^2} \leq \frac{\eta G}{(t+1)^2},$$

where $\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$.

Notes on References

The formulation of the distributed state estimation problem is adapted from [Shorinwa et al., 2020]. The distributed energy resource coordination problem is adapted from [Yang et al., 2019]. The distributed routing control problem is adapted from [Nisan et al., 2007].

The Metropolis weight matrix was adapted from the Metropolis algorithm of Markov chain Monte Carlo [Metropolis et al., 1953, Boyd et al., 2004]. A variant of the Metropolis weight

matrix is the *lazy Metropolis weight matrix*, which is given by

$$W_{ij} = \begin{cases} \frac{1}{2} + \frac{1}{2\max\{\deg(i), \deg(j)\}}, & \{i,j\} \in \mathcal{E}, \\ 0, & i \neq j \text{ and } \{i,j\} \notin \mathcal{E}, \\ 1 - \sum_{k \neq i} W_{ik}, & i = j. \end{cases}$$

The paper [Olshevsky, 2017] shows that

$$\left\| W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} \right\|_2 \leq 1 - \frac{1}{71N^2}$$

for the lazy Metropolis weight matrix, regardless of the topology of the graph. As a result, the iteration complexity of distributed averaging with the lazy Metropolis weights can be bounded by

$$O\left(N^2 \ln \frac{1}{\epsilon}\right).$$

To improve the dependence of the complexity on the number of agents N, [Olshevsky, 2017] proposes a distributed averaging protocol whose complexity can be bounded by

$$O\left(N \ln \frac{1}{\epsilon}\right).$$

The method proposed in [Olshevsky, 2017] turns out to be closedly related to applying accelerated gradient descent method to solving the optimization problem

$$\min_{x \in \mathbb{R}^N} \quad \frac{1}{2} x^\mathsf{T}(I - W)x,$$

(see also Appendix 2.A), which has been further studied in [Bu et al., 2018, Esteki et al., 2022]. There is also another approach called *Chebyshev acceleration* [Scaman et al., 2017] for accelerating the consensus method for distributed averaging.

The method for finding the optimal weight matrix that minimizes the per-step convergence factor was first proposed in [Xiao and Boyd, 2004].

For more details and some historical notes on the ratio consensus method for distributed averaging, we refer to [Hadjicostis et al., 2018]. This article is also an excellent reference for the basic consensus method and its various extensions for distributed averaging. Another excellent reference for distributed averaging and its extensions is the book [Bullo, 2022].

2.A Accelerated Consensus for Distributed Averaging

As Exercise 2.9 shows, when the weight matrix W is further chosen to be symmetric and positive semidefinite (which is the case if W is chosen to be the lazy Metropolis weight matrix), then the consensus method for distributed averaging can be derived by applying gradient descent to the optimization problem

$$\min_{z \in \mathbb{R}^N} \quad \frac{1}{2} z^\mathsf{T}(I - W)z, \tag{2.11}$$

and the convergence guarantee can be obtained by applying the convergence result of gradient descent for smooth and strongly convex objective functions. This suggests that, if we apply Nesterov's accelerated gradient descent to (2.11), we may derive a faster consensus method for distributed averaging. For simplicity, we only consider the case where each agent's associated quantity is a scalar. The following materials in this section are mostly adopted from [Olshevsky, 2017].

The version of Nesterov's accelerated gradient descent we shall apply is the following:

$$\begin{aligned} z(t+1) &= y(t) - \frac{1}{L}\nabla f(y(t)), \\ y(t+1) &= z(t+1) + \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}(z(t+1) - z(t)), \end{aligned}$$

where we initialize with $y(0) = z(0)$. Note that, while the problem (2.11) is not strongly convex, we still apply a strongly convex version of Nesterov's accelerated gradient descent. We set $L = 1$ since $\|I - W\|_2 \leq 1$. For κ, we shall set

$$\kappa \leq 1 - \sigma, \qquad \text{where} \quad \sigma = \left\| W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} \right\|_2.$$

By plugging in $f(z) = \frac{1}{2}z^\mathsf{T}(I - W)z$, we get

$$\begin{aligned} z(t+1) &= Wy(t), \\ y(t+1) &= z(t+1) + \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}(z(t+1) - z(t)), \end{aligned} \tag{2.12}$$

where we initialize with $y(0) = z(0)$. It's not hard to see that the above iterations can be implemented in a distributed manner as long as W is compatible with the topology of the network.

We first note that $\mathbf{1}^\mathsf{T}z(t)$ remains constant for the iterations (2.12). To study its convergence, we let $0 < \lambda_1 \leq \cdots \leq \lambda_{N-1} \leq 1$ be the positive eigenvalues of $I - W$, and let u_i be an eigenvector of $I - W$ with eigenvalue λ_i for each $i = 1, \ldots, N-1$ (note that $\lambda_1 = 1 - \sigma \geq \kappa$), such that $\{\mathbf{1}/\sqrt{N}, u_1, \ldots, u_{N-1}\}$ forms an orthonormal basis of $\mathbb{R}^N$. Define

$$\tilde{z}_i(t) = u_i^\mathsf{T}z(t), \qquad \tilde{y}_i(t) = u_i^\mathsf{T}y(t), \qquad i = 1, \ldots, N-1.$$

Then it can be verified that, for each $i = 1, \ldots, N-1$,

$$\begin{aligned} \tilde{z}_i(t+1) &= \tilde{y}_i(t) - \lambda_i\tilde{y}_i(t), \\ \tilde{y}_i(t+1) &= \tilde{z}_i(t+1) + \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}(\tilde{z}_i(t+1) - \tilde{z}_i(t)), \end{aligned}$$

and $\tilde{y}_i(0) = \tilde{z}_i(0)$. We can immediately recognize that the above iterations are just Nesterov's accelerated gradient descent applied to the problem

$$\min_{\tilde{z}_i \in \mathbb{R}} \quad \frac{\lambda_i}{2}\tilde{z}_i^2,$$

in which we interpret the objective function to be κ-strongly convex and 1-smooth. Therefore we can apply the convergence result of Nesterov's accelerated gradient descent (the second part of Theorem 1.8) and obtain the following convergence guarantee:

Theorem 2.5. *Suppose $W \in \mathbb{R}^{N\times N}$ is real symmetric and positive semidefinite, and satisfies $W\mathbf{1} = \mathbf{1}$ and $\sigma = \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 < 1$. Then the sequence generated by (2.12) satisfies*

$$\left\|z(t) - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} z(0)\right\|^2 \le 2\left(1-\sqrt{\kappa}\right)^t \left\|z(0) - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} z(0)\right\|^2.$$

Proof. By the second part of Theorem 1.8, we have

$$\frac{\lambda_i}{2}\tilde{z}_i(t)^2 \le \left(1-\sqrt{\kappa}\right)^t \left(\frac{\lambda_i}{2}\tilde{z}_i(0)^2 + \frac{\kappa}{2}\tilde{z}_i(0)^2\right) \le \left(1-\sqrt{\kappa}\right)^t \cdot \lambda_i \tilde{z}_i(0)^2,$$

where we used the fact that $\kappa \le \lambda_i$. Therefore, by denoting $\tilde{z}(t) = (\tilde{z}_1(t), \ldots, \tilde{z}_{N-1}(t))$, we get

$$\|\tilde{z}(t)\|^2 = \sum_{i=1}^{N-1} \tilde{z}_i(t)^2 \le 2\left(1-\sqrt{\kappa}\right)^t \sum_{i=1}^{N-1} \tilde{z}_i(0)^2 = 2\left(1-\sqrt{\kappa}\right)^t \|\tilde{z}(0)\|^2.$$

The final result follows by noting that $\|\tilde{z}(t)\| = \left\|z(t) - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} z(t)\right\|$. □

Remark 2.3. Note that, in order to achieve

$$\frac{1}{N}\sum_{i=1}^{N}(z_i(t) - \bar{z})^2 \le \epsilon^2, \qquad \bar{z} \coloneqq \frac{1}{N}\sum_{i=1}^{N} z_i(0),$$

the number of iterations T needed for the accelerated consensus method (2.12) satisfies

$$T = O\left(\frac{1}{\sqrt{1-\sigma}}\ln\frac{1}{\epsilon}\right)$$

if we choose $\kappa = 1 - \sigma$. Compared to the communication complexity bound for the vanilla consensus method

$$O\left(\frac{\ln(1/\epsilon)}{\ln(1/\sigma)}\right) = O\left(\frac{1}{1-\sigma}\ln\frac{1}{\epsilon}\right)$$

where we assume σ is sufficiently close to 1, we see that the accelerated consensus method has better scalability as σ approaches 1.

Remark 2.4. The accelerated consensus method (2.12) requires that all agents know the quantity $1 - \sigma$ and set a uniform κ, which might be difficult to achieve. On the other hand, [Olshevsky, 2017] derives the following bound when W is chosen to be the lazy Metropolis weight matrix:

$$1 - \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 \ge \frac{1}{71N^2},$$

which can be used to implement (2.12) when the total number of agents N is known to all agents. The resulting communication complexity will be $O(N\ln(1/\epsilon))$.

2.B Proof of Theorem 2.4

Denote

$$\mathsf{X}(t) = \begin{bmatrix} x_1(t)^\mathsf{T} \\ \vdots \\ x_N(t)^\mathsf{T} \end{bmatrix}, \qquad \bar{x}(t) = \frac{1}{N}\mathsf{X}(t)^\mathsf{T}\mathbf{1},$$

$$\mathsf{E}(t) = \mathsf{X}(t) - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\mathsf{X}(t), \qquad \varepsilon(t) = \frac{1}{N}\sum_{i=1}^{N}(\nabla f_i(x_i(t)) - \nabla f_i(\bar{x}(t))).$$

Then it can be checked that

$$\begin{aligned}
\bar{x}(t+1) &= \bar{x}(t) - \eta(\nabla f(\bar{x}(t)) + \varepsilon(t)),\\
\mathsf{E}(t+1) &= \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)^{K_{t+1}}\left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)\begin{bmatrix}(x_1(t) - \eta\nabla f_1(x_1(t)))^\mathsf{T}\\ \vdots \\ (x_N(t) - \eta\nabla f_N(x_N(t)))^\mathsf{T}\end{bmatrix}.
\end{aligned}$$

Note that $\|\varepsilon(t)\|$ can be bounded by

$$\begin{aligned}
\|\varepsilon(t)\|^2 &\le \frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x_i(t)) - \nabla f_i(\bar{x}(t))\|^2\\
&\le \frac{1}{N}\sum_{i=1}^{N}L^2\|x_i(t) - \bar{x}(t)\|^2 = \frac{L^2}{N}\|\mathsf{E}(t)\|_F^2.
\end{aligned}$$

To bound $\|\mathsf{E}(t+1)\|$, we first note that

$$\begin{aligned}
&\left\|\nabla f_i(x_i(t)) - \frac{1}{N}\sum_{j=1}^{N}\nabla f_j(x_j(t))\right\|\\
&\le \|\nabla f_i(x_i(t)) - \nabla f_i(\bar{x}(t))\| + \|\nabla f_i(\bar{x}(t)) - \nabla f(\bar{x}(t))\| + \left\|\frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\bar{x}(t)) - \frac{1}{N}\sum_{j=1}^{N}\nabla f_j(x_j(t))\right\|\\
&\le L\|x_i(t) - \bar{x}(t)\| + G + \frac{1}{N}\sum_{j=1}^{N}L\|x_j(t) - \bar{x}(t)\|,
\end{aligned}$$

which implies

$$\begin{aligned}
&\left\|\left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)\begin{bmatrix}(x_1(t) - \eta\nabla f_1(x_1(t)))^\mathsf{T}\\ \vdots \\ (x_N(t) - \eta\nabla f_N(x_N(t)))^\mathsf{T}\end{bmatrix}\right\|_F^2\\
&= \sum_{i=1}^{N}\left\|x_i(t) - \bar{x}(t) - \eta\left(\nabla f_i(x_i(t)) - \frac{1}{N}\sum_{j=1}^{N}\nabla f_j(x_j(t))\right)\right\|^2\\
&\le \sum_{i=1}^{N}\left(\|x_i(t) - \bar{x}(t)\| + \eta L\|x_i(t) - \bar{x}(t)\| + \eta G + \frac{\eta L}{N}\sum_{j=1}^{N}\|x_j(t) - \bar{x}(t)\|\right)^2\\
&= \left\|\begin{bmatrix}1 + \frac{1+N}{N}\eta L & \frac{1}{N}\eta L & \cdots & \frac{1}{N}\eta L\\ \frac{1}{N}\eta L & 1 + \frac{1+N}{N}\eta L & \cdots & \frac{1}{N}\eta L\\ \vdots & \vdots & \ddots & \vdots\\ \frac{1}{N}\eta L & \frac{1}{N}\eta L & \cdots & 1 + \frac{1+N}{N}\eta L\end{bmatrix}\begin{bmatrix}\|x_1(t) - \bar{x}(t)\|\\ \vdots \\ \|x_N(t) - \bar{x}(t)\|\end{bmatrix} + \eta G\mathbf{1}\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\left\| \begin{bmatrix} 1+\frac{1+N}{N}\eta L & \frac{1}{N}\eta L & \cdots & \frac{1}{N}\eta L \\ \frac{1}{N}\eta L & 1+\frac{1+N}{N}\eta L & \cdots & \frac{1}{N}\eta L \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N}\eta L & \frac{1}{N}\eta L & \cdots & 1+\frac{1+N}{N}\eta L \end{bmatrix} \right\|_2 \left\| \begin{bmatrix} \|x_1(t)-\bar{x}(t)\| \\ \vdots \\ \|x_N(t)-\bar{x}(t)\| \end{bmatrix} \right\| + \eta\sqrt{N}G \right)^2 \\
&\leq \left((1+2\eta L)\|\mathsf{E}(t)\|_F + \eta\sqrt{N}G \right)^2,
\end{aligned}
$$

where in the last step we used the Gershgorin circle theorem to bound the spectral norm of the matrix. Consequently,

$$
\begin{aligned}
\|\mathsf{E}(t+1)\|_F &\leq \sigma^{K_{t+1}}(1+2\eta L)\|\mathsf{E}(t)\|_F + \sigma^{K_{t+1}}\eta\sqrt{N}G \\
&\leq 3\sigma^{K_{t+1}}\|\mathsf{E}(t)\|_F + \sigma^{K_{t+1}}\eta\sqrt{N}G,
\end{aligned}
$$

where we used $\eta L \leq 1$. Since we initialize with $x_{i,0} = 0$, we have $\mathsf{E}(0) = 0$. Furthermore, by our choice of K_t, we have

$$
\sigma^{K_{t+1}} \leq \frac{1}{4((t+1)^2+3)} \leq \frac{(t+1)^2}{((t+1)^2+3)(t+2)^2}.
$$

Therefore, as long as $\|\mathsf{E}(t)\|_F \leq \eta\sqrt{N}G/(t+1)^2$, we can get

$$
\begin{aligned}
\|\mathsf{E}(t+1)\|_F &\leq \eta\sqrt{N}\sigma^{K_{t+1}}\left(\frac{3G}{(t+1)^2} + G\right) \\
&\leq \eta\sqrt{N}\frac{(t+1)^2}{((t+1)^2+3)(t+2)^2} \cdot \frac{3G + G(t+1)^2}{(t+1)^2} = \eta\sqrt{N}\frac{G}{(t+2)^2}.
\end{aligned}
$$

By mathematical induction, we obtain

$$
\|\mathsf{E}(t)\|_F \leq \eta\sqrt{N}\frac{G}{(t+1)^2},
$$

which leads to

$$
\|\varepsilon(t)\| \leq \eta L\frac{G}{(t+1)^2}.
$$

We now employ the following proposition for studying the evolution of $\bar{x}(t)$:

Proposition 2.4 (A simplifed version of [Schmidt et al., 2011, Proposition 1]). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth. Consider the following iterations:*

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

where $\eta \in (0, 1/L]$. Then for all $t \geq 1$, we have

$$
f\left(\frac{1}{t}\sum_{\tau=1}^{t} x_\tau\right) - f(x^*) \leq \frac{1}{2\eta t}\left(\|x_0 - x^*\| + A_t\right)^2,
$$

where $A_t = \sum_{\tau=0}^{t-1}\eta\|e_\tau\|$.

In order to apply Proposition 2.4, we calculate

$$\sum_{\tau=0}^{t-1} \eta \|\varepsilon(\tau)\| \leq \eta^2 LG \sum_{\tau=0}^{t-1} \frac{1}{(\tau+1)^2} \leq 2\eta^2 LG.$$

Then by Proposition 2.4, we get

$$f\left(\frac{1}{t}\sum_{\tau=1}^{t} \bar{x}(\tau)\right) - f(x^*) \leq \frac{1}{2\eta t}\left(\|x^*\| + 2\eta^2 LG\right)^2.$$

The bound on the consensus error can be obtained by multiplying $\|\mathsf{E}(t)\|_F$ with $1/\sqrt{N}$.

Bibliography

[Boyd et al., 2004] Boyd, S., Diaconis, P., and Xiao, L. (2004). Fastest mixing Markov chain on a graph. *SIAM Review*, 46(4):667–689.

[Bu et al., 2018] Bu, J., Fazel, M., and Mesbahi, M. (2018). Accelerated consensus with linear rate of convergence. In *2018 Annual American Control Conference*, pages 4931–4936.

[Bullo, 2022] Bullo, F. (2022). *Lectures on Network Systems*. Kindle Direct Publishing, 1.6 edition.

[Esteki et al., 2022] Esteki, A.-S., Moradian, H., and Kia, S. S. (2022). The fastest linearly converging discrete-time average consensus using buffered information. *arXiv preprint arXiv:2206.09916*.

[Gharesifard and Cortés, 2010] Gharesifard, B. and Cortés, J. (2010). When does a digraph admit a doubly stochastic adjacency matrix? In *Proceedings of the 2010 American Control Conference*, pages 2440–2445.

[Hadjicostis et al., 2018] Hadjicostis, C. N., Domínguez-García, A. D., and Charalambous, T. (2018). Distributed averaging and balancing in network systems: with applications to coordination and control. *Foundations and Trends® in Systems and Control*, 5(2-3):99–292.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

[Nisan et al., 2007] Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V., editors (2007). *Algorithmic Game Theory*. Cambridge University Press.

[Olshevsky, 2017] Olshevsky, A. (2017). Linear time average consensus and distributed optimization on fixed graphs. *SIAM Journal on Control and Optimization*, 55(6):3990–4014.

[Scaman et al., 2017] Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. (2017). Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3027–3036.

[Schmidt et al., 2011] Schmidt, M., Roux, N., and Bach, F. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

[Shorinwa et al., 2020] Shorinwa, O., Yu, J., Halsted, T., Koufos, A., and Schwager, M. (2020). Distributed multi-target tracking for autonomous vehicle fleets. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3495–3501. IEEE.

[Xiao and Boyd, 2004] Xiao, L. and Boyd, S. (2004). Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78.

[Yang et al., 2019] Yang, T., Yi, X., Wu, J., Yuan, Y., Wu, D., Meng, Z., Hong, Y., Wang, H., Lin, Z., and Johansson, K. H. (2019). A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305.

Chapter 3

Decentralized Gradient Descent

3.1 The Algorithm

Consider the consensus optimization problem

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

where each $f_i : \mathbb{R}^d \to \mathbb{R}$ denotes the local cost function of agent i, and f is the global objective function. We make the following assumptions:

1. The communication network is static and bi-directional, whose topology is given by an undirected graph $\mathcal{G} = (\{1, \ldots, N\}, \mathcal{E})$.
2. Each local cost function f_i is differentiable.

Let $W \in \mathbb{R}^{N \times N}$ be a weight matrix that is compatible with the graph $\mathcal{G}$. By the results from the last chapter, we can choose W to satisfy the following condition, which will be assumed throughout this chapter:

$$\sigma \coloneqq \left\| W - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right\|_2 < 1.$$

The decentralized gradient descent (DGD) method is given by

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

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

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

This variant is also called the *diffusion* method.

Just as in the previous chapter, we introduce the notation

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

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

$$F(\mathsf{X}) = \sum_{i=1}^{N} f_i(x_i), \quad \text{where } \mathsf{X} = \begin{bmatrix} — \, x_1^\mathsf{T} \, — \\ \vdots \\ — \, x_N^\mathsf{T} \, — \end{bmatrix}. \tag{3.1}$$

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

$$\nabla F(\mathsf{X}) = \begin{bmatrix} \frac{\partial F(\mathsf{X})}{\partial \mathsf{X}_{1,1}} & \cdots & \frac{\partial F(\mathsf{X})}{\partial \mathsf{X}_{1,d}} \\ \vdots & \ddots & \vdots \\ \frac{\partial F(\mathsf{X})}{\partial \mathsf{X}_{N,1}} & \cdots & \frac{\partial F(\mathsf{X})}{\partial \mathsf{X}_{N,d}} \end{bmatrix} = \begin{bmatrix} — \, \nabla f_1(x_1)^\mathsf{T} \, — \\ \vdots \\ — \, \nabla f_N(x_N)^\mathsf{T} \, — \end{bmatrix}, \qquad \text{where} \mathsf{X} = \begin{bmatrix} — \, x_1^\mathsf{T} \, — \\ \vdots \\ — \, x_N^\mathsf{T} \, — \end{bmatrix}.$$

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

$$\mathsf{X}(t+1) = W\mathsf{X}(t) - \eta_t \nabla F(\mathsf{X}(t))$$

and

$$\mathsf{X}(t+1) = W(\mathsf{X}(t) - \eta_t \nabla F(\mathsf{X}(t))),$$

respectively. In this chapter, 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.

Exercise 3.1. Suppose each f_i is convex and L-smooth, and let $\mathbb{R}^{N \times d}$ be equipped with the inner product

$$\langle \mathsf{X}, \mathsf{Y} \rangle \coloneqq \operatorname{tr}(\mathsf{X}^\mathsf{T} \mathsf{Y}).$$

Show that the corresponding F defined by (3.1) is convex and L-smooth, i.e.,

$$0 \le F(\mathsf{Y}) - F(\mathsf{X}) - \langle \nabla F(\mathsf{X}), \mathsf{Y} - \mathsf{X} \rangle \le \frac{L}{2} \|\mathsf{Y} - \mathsf{X}\|_F^2$$

for any $\mathsf{X}, \mathsf{Y} \in \mathbb{R}^{N \times d}$. □

3.2 Useful Observations and Tools for Convergence Analysis

We let $\bar{x}(t)$ denote the averaged iterate among all agents, defined by

$$\bar{x}(t) = \frac{1}{N} \sum_{i=1}^{N} x_i(t).$$

It's not hard to verify that

$$\bar{x}(t) = \frac{1}{N}\left(\mathbf{1}^\mathsf{T}\mathsf{X}(t)\right)^\mathsf{T}$$

and that

$$\begin{aligned}\bar{x}(t+1) &= \bar{x}(t) - \eta_t \frac{1}{N}(\mathbf{1}^\mathsf{T}\nabla F(\mathsf{X}(t)))^\mathsf{T} \\ &= \bar{x}(t) - \eta_t \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i(t)).\end{aligned} \tag{3.2}$$

Next, we let $\mathsf{E}(t)$ denote the matrix of consensus errors defined by

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

Then it can be checked that

$$\mathsf{E}(t+1) = \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)\mathsf{E}(t) - \eta_t\left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)\nabla F(\mathsf{X}(t)). \tag{3.3}$$

As we shall see, our convergence analysis of DGD roughly consists of the following ingredients:

1. Analyzing how the consensus error $\mathsf{E}(t)$ changes as t increases. Ideally, we wish to establish $\mathsf{E}(t) \to 0$ as $t \to \infty$. However, as long as $\bar{x}(t)$ can eventually provide a good solution to the minimization of $f(x)$, it may be okay if we only require $\mathsf{E}(t)$ to remain bounded, as we can run additional consensus steps after completing the DGD iterations.
2. Analyzing how the averaged iterate $\bar{x}(t)$ evolves with time. Particularly, we will be interested in whether $f(\bar{x}(t))$ can approach the optimal value f^* as t increases.

The Consensus Error $\mathsf{E}(t)$

Analysis of the consensus error will be mainly based on the equality (3.3). Particularly, since we have assumed that

$$\sigma = \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 < 1,$$

By denoting

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

we directly have

$$\|\mathsf{E}(t+1)\|_F \le \sigma\|\mathsf{E}(t)\|_F + \eta_t\|\Delta(t)\|_F. \tag{3.4}$$

This inequality leads to the following result:

Lemma 3.1. *Suppose there exists $\delta > 0$ such that $\|\Delta(t)\| \le \delta$ for all t. Let $\eta_0 \ge \eta_1 \ge \eta_2 \ge \cdots$ be a sequence of non-increasing step sizes. Then*

$$\|\mathsf{E}(t)\|_F \le \sigma^t\|\mathsf{E}(0)\|_F + \delta\sum_{\tau=0}^{t-1}\sigma^{t-1-\tau}\eta_\tau, \tag{3.5}$$

and

$$\sum_{\tau=0}^{t-1}\eta_\tau\|\mathsf{E}(\tau)\|_F^2 \le \frac{2\eta_0\|\mathsf{E}(0)\|_F^2}{1-\sigma^2} + \frac{4\delta^2}{(1-\sigma^2)^2}\sum_{\tau=0}^{t-2}\eta_\tau^3 \tag{3.6}$$

Proof. By mathematical induction, we can derive from (3.4) that

$$\|\mathsf{E}(t)\|_F \le \sigma^t\|\mathsf{E}(0)\|_F + \sum_{\tau=0}^{t-1}\sigma^{t-1-\tau}\eta_\tau\|\Delta(\tau)\|_F, \tag{3.7}$$

By plugging in $\|\Delta(t)\|_F \le \delta$, we get the inequality (3.5).

To show (3.6), we use (3.4) to obtain

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

where we used the inequality $\|u+v\|^2 \le (1+\epsilon)\|u\|^2+(1+1/\epsilon)\|v\|^2$ for any $\epsilon > 0$. By mathematical induction, we get

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

Consequently,

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

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

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

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

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

Therefore

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

which completes the proof. □

Exercise 3.2. Let $f : (0, +\infty) \to [0, +\infty)$ be a strictly decreasing function, and let $t \in \mathbb{N}\backslash\{0\}$ be arbitrary.

1. Show that
$$\sum_{\tau=1}^{t} f(\tau) > \int_{1}^{t+1} f(x)\,\mathrm{d}x.$$
2. Suppose further that f is also convex, show that
$$\int_{a}^{b} f(x)\,\mathrm{d}x \geq (b-a) f\left(\frac{a+b}{2}\right)$$
for any $0 < a \leq b$. Then use the above inequality to prove that
$$\sum_{\tau=1}^{t} f(\tau) \leq \int_{1/2}^{t+1/2} f(x)\,\mathrm{d}x.$$
3. Let $\beta \in (0, 1]$ be arbitrary. Use the above results to derive lower and upper bounds for
$$\sum_{\tau=1}^{t} \frac{1}{\tau^{\beta}}.$$

□

Exercise 3.3. Let $\sigma \in (0, 1)$ and $\beta > 0$ be arbitrary. Show that
$$\lim_{t\to\infty} (1-\sigma) t^{\beta} \sum_{\tau=1}^{t} \frac{\sigma^{t-\tau}}{\tau^{\beta}} = 1.$$
(Hint: You may consider using the Stolz–Cesàro theorem: Given two sequences $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ of real numbers, suppose that $(b_n)_{n\in\mathbb{N}}$ is strictly increasing and $\lim_{n\to\infty} b_n = +\infty$, and that
$$\lim_{n\to\infty} \frac{a_{n+1} - a_n}{b_{n+1} - b_n} = r$$
for some $r \in \mathbb{R}$. Then $\lim_{n\to\infty} a_n / b_n = r$.)

□

Exercise 3.4. Let $\eta_0 \geq \eta_1 \geq \eta_2 \geq \cdots$ be a non-increasing sequence of step sizes. Suppose $\mathsf{E}(t) \in \mathbb{R}^{N\times d}$ satisfies
$$\|\mathsf{E}(t+1)\|_F \leq \sigma \|\mathsf{E}(t)\|_F + \eta_t \delta,$$
where $\sigma \in (0, 1)$, $\delta > 0$, and $\eta_t > 0$ for all t. Show that
$$\sum_{\tau=0}^{t-1} \eta_\tau \|\mathsf{E}(\tau)\|_F \leq \frac{\eta_0 \|\mathsf{E}(0)\|_F}{1-\sigma} + \frac{\delta}{1-\sigma} \sum_{\tau=0}^{t-2} \eta_\tau^2.$$

□

The Averaged Iterate

We now study the averaged iterate $\bar{x}(t)$. We first make the following assumptions on the local cost functions:

Assumption 3.1. 1. Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex, G-Lipschitz continuous and L-smooth.

2. There exists $x^* \in \mathbb{R}^d$ such that $f(x^*) = \inf_{x \in \mathbb{R}^d} f(x)$.

We introduce the notation

$$\bar{g}(t) = \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_i(t)) = \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \nabla F(\mathsf{X}(t)).$$

It can be seen that $\bar{x}(t)$ satisfies the following iteration

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

We begin our analysis by noting that

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

which leads to

$$\frac{1}{2}\|\bar{x}(t+1) - x^*\|^2 = \eta_t \langle \bar{g}(t), x^* - \bar{x}(t+1) \rangle + \frac{1}{2}\|\bar{x}(t) - x^*\|^2 - \frac{1}{2}\|\bar{x}(t+1) - \bar{x}(t)\|^2. \tag{3.8}$$

Just like in the analysis of centralized GD, we try to associate the inner product $\langle \bar{g}(t), x^* - \bar{x}(t+1) \rangle$ with the objective values $f(\bar{x}(t+1))$ and $f(x^*)$. Note that

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

For the first term, we have

$$\begin{aligned} \langle \bar{g}(t), x^* - \bar{x}(t) \rangle &= \frac{1}{N} \sum_{i=1}^{N} \langle \nabla f_i(x_i(t)), x^* - \bar{x}(t) \rangle \\ &= \frac{1}{N} \sum_{i=1}^{N} \langle \nabla f_i(x_i(t)), x^* - x_i(t) \rangle + \frac{1}{N} \sum_{i=1}^{N} \langle \nabla f_i(x_i(t)), x_i(t) - \bar{x}(t) \rangle \\ &\leq \frac{1}{N} \sum_{i=1}^{N} (f_i(x^*) - f_i(x_i(t))) + \frac{1}{N} \sum_{i=1}^{N} \langle \nabla f_i(x_i(t)), x_i(t) - \bar{x}(t) \rangle \\ &= f(x^*) - \hat{f}(\mathsf{X}(t)), \end{aligned} \tag{3.9}$$

where we employ the convexity of each f_i in the first inequality, and denote

$$\hat{f}(\mathsf{X}) \coloneqq \frac{1}{N} \sum_{i=1}^{N} (f_i(x_i) + \langle \nabla f_i(x_i), \bar{x} - x_i \rangle) = \frac{1}{N} \left(F(\mathsf{X}) + \operatorname{tr} \left[\nabla F(\mathsf{X}) \left(\frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \mathsf{X} - \mathsf{X} \right)^\mathsf{T} \right] \right).$$

To bound the second term, we note that the L-smoothness of each f_i implies

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

By taking the average over $i = 1, \dots, N$, we get

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

Therefore

$$\langle \bar{g}(t), \bar{x}(t) - \bar{x}(t+1)\rangle \le \hat{f}(\mathsf{X}(t)) - f(\bar{x}(t+1)) + \frac{L}{2}\|\bar{x}(t+1) - \bar{x}(t)\|^2 + \frac{L}{2N}\|\mathsf{E}(t)\|_F^2.$$

By adding the two bounds, we get

$$\langle \bar{g}(t), x^* - \bar{x}(t+1)\rangle \le f(x^*) - f(\bar{x}(t+1)) + \frac{L}{2}\|\bar{x}(t+1) - \bar{x}(t)\|^2 + \frac{L}{2N}\|\mathsf{E}(t)\|_F^2,$$

and after plugging this inequality into (3.8), we get

$$\begin{aligned} &\eta_t f(\bar{x}(t+1)) + \frac{1}{2}\|\bar{x}(t+1) - x^*\|^2 \\ &\le \eta_t f(x^*) + \frac{1}{2}\|\bar{x}(t) - x^*\|^2 + \frac{\eta_t L}{2N}\|\mathsf{E}(t)\|_F^2 + \frac{\eta_t L - 1}{2}\|\bar{x}(t+1) - \bar{x}(t)\|^2. \end{aligned}$$

By taking the telescoping sum, we can obtain the following lemma:

Lemma 3.2. *Suppose Assumption 3.1 holds, and $\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$ satisfies*

$$\bar{x}(t+1) = \bar{x}(t) - \eta_t \cdot \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_i(t)), \tag{3.10}$$

where $\eta_t > 0$ for each $t \in \mathbb{N}$. Then

$$\begin{aligned} &\frac{\sum_{\tau=1}^{t} \eta_{\tau-1}(f(\bar{x}(\tau)) - f(x^*))}{\sum_{\tau=1}^{t} \eta_{\tau-1}} \\ &\le \frac{\|\bar{x}(0) - x^*\|^2}{2\sum_{\tau=0}^{t-1} \eta_\tau} + \frac{L}{2N} \cdot \frac{\sum_{\tau=0}^{t-1} \eta_\tau \|\mathsf{E}(\tau)\|_F^2}{\sum_{\tau=0}^{t-1} \eta_\tau} + \frac{\sum_{\tau=0}^{t-1} (\eta_\tau L - 1)\|\bar{x}(\tau+1) - \bar{x}(\tau)\|^2}{2\sum_{\tau=0}^{t-1} \eta_\tau}, \end{aligned}$$

where

$$\mathsf{E}(t) = \begin{bmatrix} (x_1(t) - \bar{x}(t))^\mathsf{T} \\ \vdots \\ (x_N(t) - \bar{x}(t))^\mathsf{T} \end{bmatrix}.$$

Remark 3.1. Note that the derivation of Lemma 3.2 requires the local variable iterates $x_i(t)$ to only satisfy (3.10) but does not assume any detailed local update rule. As a result, Lemma 3.2 may also be used for the analysis of other distributed optimization algorithms as long as (3.10) (and other technical conditions on the local cost functions) are satisfied. Particularly, Lemma 3.2 will be employed when analyzing the gradient tracking algorithm, which employs more complicated local update rules but still satisfies (3.10).

Remark 3.2. The above derivations adopt the routine that we mainly focus on the evolution of $\bar{x}(t)$ and $f(\bar{x}(t)) - f(x^*)$, and whenever a quantity involving individual $x_i(t)$ cannot be cancelled out, we approximate it by $\bar{x}(t)$ and bound the error via $\|\mathsf{E}(t)\|_F$. In the literature, there is also another approach that focuses on the evolution of $\frac{1}{N}F(\mathsf{X}(t)) - f(x^*) = \frac{1}{N}\sum_{i=1}^N (f_i(x_i(t)) - f(x^*))$ and only in the very end derives bounds on $f(\bar{x}(t)) - f(x^*)$ via $\|\mathsf{E}(t)\|_F$. Here we adopt the former approach as it seems more similar to the analysis of the centralized GD, but the latter approach may be useful for some more complicated settings (e.g., constrained problems, composite objective functions, etc.).

3.3 Convergence Analysis: The Convex and Smooth Case

By combining Lemmas 3.1 and 3.2, we can now derive convergence results for DGD.

Theorem 3.1. *Suppose each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex, G-Lipschitz and L-smooth, and suppose $f(x^*) = \inf_{x\in\mathbb{R}^d} f(x)$ for some $x^* \in \mathbb{R}^d$. Let $\eta_0 \geq \eta_1 \geq \eta_2 \geq \cdots$ be a sequence of non-increasing positive step sizes satisfying $\eta_0 L \leq 1$. Denote*

$$E_0 = \sqrt{\frac{1}{N}\sum_{i=1}^N \|x_i(0) - \bar{x}(0)\|^2}.$$

Then

$$\frac{\sum_{\tau=1}^t \eta_{\tau-1}(f(\bar{x}(\tau)) - f(x^*))}{\sum_{\tau=1}^t \eta_{\tau-1}} \leq \frac{\|\bar{x}(0) - x^*\|^2}{2\sum_{\tau=0}^{t-1}\eta_\tau} + \frac{\eta_0 L E_0^2}{(1-\sigma^2)\sum_{\tau=0}^{t-1}\eta_\tau} + \frac{2LG^2\sum_{\tau=0}^{t-2}\eta_\tau^3}{(1-\sigma^2)^2\sum_{\tau=0}^{t-1}\eta_\tau}.$$

Moreover,

$$\frac{1}{N}\sum_{i=1}^N \|x_i(t) - \bar{x}(t)\|^2 \leq \left(\sigma^t E_0 + G\sum_{\tau=0}^{t-1}\sigma^{t-1-\tau}\eta_\tau\right)^2.$$

Proof. Since $\eta_t L \leq \eta_0 L \leq 1$ for all t, from Lemma 3.2, we get

$$\frac{\sum_{\tau=1}^t \eta_{\tau-1}(f(\bar{x}(\tau)) - f(x^*))}{\sum_{\tau=1}^t \eta_{\tau-1}} \leq \frac{\|\bar{x}(0) - x^*\|^2}{2\sum_{\tau=0}^{t-1}\eta_\tau} + \frac{L}{2N}\cdot\frac{\sum_{\tau=0}^{t-1}\eta_\tau\|\mathsf{E}(\tau)\|_F^2}{\sum_{\tau=0}^{t-1}\eta_\tau}. \tag{3.11}$$

Then, since each f_i is G-Lipschitz continuous, by Lemma 1.2 and Proposition 1.2, we have $\|\nabla f_i(x)\| \leq G$ for all $x \in \mathbb{R}^d$ and all i. Therefore

$$\begin{aligned}\|\Delta(t)\|_F^2 &= \left\| \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} \right) \nabla F(\mathsf{X}(t)) \right\|_F^2 \leq \|\nabla F(\mathsf{X}(t))\|_F^2 \\ &= \sum_{i=1}^N \|\nabla f_i(x_i(t))\|^2 \leq NG^2,\end{aligned}$$

where in the first inequality we used $\left\| I - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} \right\|_2 \leq 1$. We can now apply Lemma 3.1 and obtain

$$\|\mathsf{E}(t)\|_F \leq \sigma^t \|\mathsf{E}(0)\|_F + \sqrt{N}G \sum_{\tau=0}^{t-1} \sigma^{t-1-\tau}\eta_\tau$$

and

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

The bound for the consensus error is now evident. By plugging the above bound into (3.11) and noting that $E_0 = \|\mathsf{E}(0)\|_F/\sqrt{N}$, we get the desired convergence results for the objective values. □

Corollary 3.1. *Let the conditions of Theorem 3.1 be satisfied. For simplicity, suppose every agent starts from the same initial point so that $E_0 = 0$.*

1. *Suppose we choose a constant step size $\eta_t = \eta \leq 1/L$. Then*

$$\frac{1}{t}\sum_{\tau=1}^{t} (f(\bar{x}(t)) - f(x^*)) \leq \frac{\|\bar{x}(0) - x^*\|^2}{2\eta t} + \frac{2\eta^2 LG^2}{(1-\sigma^2)^2},$$

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. *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*

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

and

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

Exercise 3.5. Prove Corollary 3.1. □

Some discussions of Theorem 3.1 and Corollary 3.1 are provided as follows:

1. The conditions of Theorem 3.1 include the Lipschitz continuity of each local cost function, which is not required for the convergence of centralized GD and also does not hold for some problem scenarios we may encounter. This assumption ensures that

$$\|\Delta(t)\|_F^2 = \sum_{i=1}^{N} \left\| \nabla f_i(x_i(t)) - \frac{1}{N} \sum_{j=1}^{N} \nabla f_j(x_j(t)) \right\|^2$$

is upper bounded throughout the DGD iterations. One may wonder whether we can establish the boundedness of $\|\Delta(t)\|_F$ automatically without imposing additional assumptions. However, as the following example shows, if we only assume convexity and L-smoothness for each local cost function, we may not able to derive a bound on the consensus error.

Example 3.1 ([Jakovetic et al., 2011])**.** Consider a 2-agent system. For an arbitrary $\theta > 0$, let the local cost functions $f_i^\theta : \mathbb{R} \to \mathbb{R}$, $i = 1, 2$ be defined by

$$f_i^\theta(x) = \frac{1}{2}(x + (-1)^i \theta)^2.$$

It's not hard to see that each f_i^θ is convex and 1-smooth. Let the weight matrix be given by

$$W = \frac{1}{4} \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}.$$

Let the initial points be $x_1(0) = x_2(0) = 0$, let the step sizes be $\eta_t = \eta_0/(t+1)^\beta$ with $\beta \in (0, 1)$ and $\eta_0 \leq 1/2$, and let $(x_i(t))_{t \geq 1}$ be generated by the DGD algorithm. Then it's not hard to see that $x_1(t) = -x_2(t) \geq 0$, and

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

Therefore the consensus error satisfies

$$\frac{1}{2} \sum_{i=1}^{2} \|x_i(t) - \bar{x}(t)\|^2 \geq \eta_t^2 \theta^2 = \frac{\eta_0^2 \theta^2}{(t+1)^{2\beta}}.$$

We see that, for each fixed t, the consensus error can be made arbitrarily large by letting θ be sufficiently large. □

The above example suggests that, even if each f_i is assumed to be convex and smooth, a bound on the consensus error may inevitably depend on some other parameter of the problem that is associated with bounding $\|\Delta(t)\|_F$. In the next section, we will present an approach to replace the condition of Lipschitz continuity with a milder (but not necessarily weaker) one.

2. When we choose a constant step size $\eta_t = \eta$, the first part of Corollary 3.1 shows that there is a $O(\eta^2/(1-\sigma)^2)$ gap that does not vanish as $t \to \infty$, suggesting that the DGD algorithm does not converge exactly to the optimal. This is also in accordance with the observation that the optimal solution $\mathsf{X}^* = \mathbf{1}x^{*\mathsf{T}}$ is not a fixed-point of the DGD iteration $\mathsf{X}(t+1) = W\mathsf{X}(t) - \eta\nabla F(\mathsf{X}(t))$.

3. When we employ a diminishing sequence of step sizes of the form $\eta_t \propto 1/(t+1)^\beta$ for $\beta \in (0,1)$, the second part of Corollary 3.1 shows that the DGD iterations will converge but with an inferior rate compared to centralized GD. Particularly, choosing $\beta = 1/3$ provides the optimal rate $O\big(\ln t/t^{2/3}\big)$, slower than the $O(1/t)$ rate achieved by centralized GD. This gap is one of the motivations for proposing gradient-tracking-based distributed optimization algorithms that can mitigate this gap.

   We note that the $O\big(\ln t/t^{2/3}\big)$ rate derived here is better than the $O(\ln t/\sqrt{t})$ rate usually found in existing literature (e.g., [Chen, 2012]). On the other hand, this rate agrees with the $\Omega(1/t^{2/3})$ lower bound on the worst-case optimality gap presented in [Jakovetic et al., 2011] up to a logarithmic factor.

3.4 Another Perspective of DGD with Constant Step Sizes

In this section, we analyze DGD with a constant step size from another perspective. Specifically, when the step size is constant, we may view the DGD iterations as the gradient descent updates for an optimization problem. This perspective allows us to replace the condition $\|\nabla f_i(x)\| \le G$ for DGD's convergence when η_t is constant.

We make the following assumptions on the local cost functions f_i and the global objective function f:

Assumption 3.2. Each local cost function $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth, and

$$f_i^\ell \coloneqq \inf_{x\in\mathbb{R}^d} f_i(x) > -\infty.$$

Note that we do not assume $\|\nabla f_i(x)\|$ to be uniformly bounded any more. Instead, we assume that each f_i is lower bounded, which seem to have wider applicability. The idea behind this "relaxation" of the condition is demonstrated by the following lemma:

Lemma 3.3. *Suppose $h : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, and has a finite lower bound $h^\ell = \inf_{x\in\mathbb{R}^d} h(x) > -\infty$. Then*

$$\|\nabla h(x)\|^2 \le 2L(h(x) - h^\ell), \qquad \forall x \in \mathbb{R}^d.$$

Proof. We have

$$h^\ell \le h\bigg(x - \frac{1}{L}\nabla h(x)\bigg) \le h(x) - \frac{1}{2L}\|\nabla h(x)\|^2,$$

which then implies the desired inequality. □

The result of Lemma 3.3 suggests that, in the DGD iterations, if we can ensure that each $f_i(x_i(t)) - f_i^\ell$ is upper bounded, then each $\|\nabla f_i(x_i(t))\|$, and consequently $\|\Delta(t)\|_F$, will be upper bounded. Since $f_i(x_i(t)) - f_i^\ell \geq 0$ for any i, we have that each $f_i(x_i(t)) - f_i^\ell$ is upper bounded if and only if

$$F(\mathsf{X}(t)) = \sum_{i=1}^{N} (f_i(x_i(t)) - f_i^\ell) + \sum_{i=1}^{N} f_i^\ell$$

is upper bounded. Recall that the DGD iterations can be written as

$$\mathsf{X}(t+1) = W\mathsf{X}(t) - \eta \nabla F(\mathsf{X}(t)).$$

If we can associate the DGD iterations with the gradient descent iterations of some optimization problem with $F(\mathsf{X})$ being part of its objective function, then we might be able to establish the upper-boundedness of $F(\mathsf{X}(t))$ since GD is a descent algorithm.

We now present how the above idea can be realized in detail. We make the following assumptions on the weight matrix:

Assumption 3.3. 1. W is real symmetric and positive semidefinite.

2. $W\mathbf{1} = \mathbf{1}$, and

$$\sigma \coloneqq \left\| W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} \right\|_2 < 1.$$

Note that the positive semidefiniteness of W does not really restrict the applicability of our theory: For any symmetric $\tilde{W} \in \mathbb{R}^{N\times N}$ satisfying $\tilde{W}\mathbf{1} = \mathbf{1}$ and $\left\| \tilde{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T} \right\|_2 < 1$, we can construct a new weight matrix $W = (\tilde{W} + I)/2$ which is positive semidefinite.

We then define the seminorm[1]

$$\|\mathsf{X}\|_{I-W} \coloneqq \sqrt{\operatorname{tr}(\mathsf{X}^\mathsf{T}(I-W)\mathsf{X})}, \qquad \forall \mathsf{X} \in \mathbb{R}^{N\times d}.$$

Note that $\|\cdot\|_{I-W}$ is only a seminorm since $\|\mathsf{X}\|_{I-W} = 0$ does not imply $\mathsf{X} = 0$. Instead, we have

Lemma 3.4. *$\|\mathsf{X}\|_{I-W}^2 = 0$ if and only if X has identical row vectors. Moreover,*

$$\left\| \mathsf{X} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\mathsf{X} \right\|_F^2 \leq \frac{1}{1-\sigma}\|\mathsf{X}\|_{I-W}^2. \tag{3.12}$$

Proof. If X has identical row vectors, then $W\mathsf{X} = \mathsf{X}$, which implies that $(I-W)\mathsf{X} = 0$ and subsequently $\|\mathsf{X}\|_{I-W} = 0$.

We then proceed to show the inequality (3.12). Let $x^{(1)}, \ldots, x^{(d)}$ be the *column* vectors of X. It can be checked that

$$\|\mathsf{X}\|_{I-W}^2 = \sum_{j=1}^{d} x^{(j)\mathsf{T}}(I-W)x^{(j)}.$$

[1] A seminorm is a function p defined on a vector space that i) takes nonnegative values, ii) satisfies $p(\lambda v) = |\lambda| \cdot p(v)$ for any scalar λ and vector v, and iii) satisfies the triangle inequality.

Then since

$$I - W = \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)(I - W)\left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right),$$

we get

$$\|\mathsf{X}\|_{I-W}^2 = \sum_{j=1}^{d}\left(x^{(j)} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x^{(j)}\right)^\mathsf{T}(I - W)\left(x^{(j)} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x^{(j)}\right).$$

Note that the vector $x^{(j)} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x^{(j)}$ is orthogonal to $\mathbf{1}$, and therefore

$$\begin{aligned}
&\left(x^{(j)} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x^{(j)}\right)^\mathsf{T}(I - W)\left(x^{(j)} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x^{(j)}\right) \\
\geq\ &\left\|x^{(j)} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x^{(j)}\right\|^2 \cdot \min_{x\in\mathbb{R}^N:\langle x,\mathbf{1}\rangle=0}\frac{x^\mathsf{T}(I-W)x}{\|x\|^2} \\
=\ &(1-\sigma)\left\|x^{(j)} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x^{(j)}\right\|^2,
\end{aligned}$$

where we used the fact that the two smallest eigenvalues of $I - W$ are 0 and $1 - \sigma$, while the eigenspace associated with the eigenvalue 0 is $\mathrm{span}\{\mathbf{1}\}$. Consequently,

$$\|\mathsf{X}\|_{I-W}^2 \geq (1-\sigma)\sum_{j=1}^{d}\left\|x^{(j)} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}x^{(j)}\right\|^2 = (1-\sigma)\left\|\mathsf{X} - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\mathsf{X}\right\|_F^2,$$

which is just (3.12).

Finally, by using the inequality (3.12), it's not hard to see that $\|\mathsf{X}\|_{I-W} = 0$ implies $\mathsf{X} = \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\mathsf{X}$, i.e., X has identical row vectors. □

Exercise 3.6. Suppose the weight matrix $W \in \mathbb{R}^{N\times N}$ satisfies Assumption 3.3. Let

$$g(\mathsf{X}) = \frac{1}{2}\|\mathsf{X}\|_{I-W}^2, \qquad \mathsf{X} \in \mathbb{R}^{N\times d}.$$

Show that

1. $\nabla g(\mathsf{X}) = (I - W)\mathsf{X}$.
2. $g(\mathsf{X})$ is 1-smooth. □

Now let us consider the following optimization problem:

$$\min_{\mathsf{X}\in\mathbb{R}^{N\times d}} \quad F(\mathsf{X}) + \frac{1}{2\eta}\|\mathsf{X}\|_{I-W}^2, \tag{3.13}$$

where $\eta > 0$ is a constant. By the result of Exercise 3.6, we see that the gradient descent iteration with a constant step size η for solving (3.13) is given by

$$\mathsf{X}(t+1) = W\mathsf{X}(t) - \eta\nabla F(\mathsf{X}(t)),$$

which is just the DGD method with a constant step size η. By the results of Exercises 3.1 and 3.6, the objective function of (3.13) is $(L+1/(2\eta))$-smooth. Therefore, as long as $\eta(L+1/(2\eta)) \le 2$, i.e., $\eta \le 3/(2L)$, we can use the descent property of GD to get

$$F(\mathsf{X}(t)) \le F(\mathsf{X}(t)) + \frac{1}{2\eta}\|\mathsf{X}(t)\|_{I-W}^2 \le F(\mathsf{X}(0)) + \frac{1}{2\eta}\|\mathsf{X}(0)\|_{I-W}^2.$$

By employing Lemma 3.3, we arrive at the following theorem:

Theorem 3.2. *Suppose Assumptions 3.3 and 3.2 hold, and suppose $x^* \in \mathbb{R}^d$ is a minimizer of $f(x)$. Let the agents share the same initial point $x_i(0) = x_0$, and denote*

$$D = \sqrt{f(x_0) - \frac{1}{N}\sum_{i=1}^N f_i^\ell}.$$

Then for the DGD iteration with constant step size $\eta \le 1/L$, we have

$$\|\Delta(t)\|_F^2 \le 2LND^2.$$

Consequently,

$$\frac{1}{t}\sum_{\tau=1}^t (f(\bar{x}(\tau)) - f(x^*)) \le \frac{\|x_0 - x^*\|^2}{2\eta t} + \frac{4\eta^2 L^2 D^2}{(1-\sigma^2)^2},$$

and

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

Exercise 3.7. Suppose the weight matrix $W \in \mathbb{R}^{N\times N}$ satisfies Assumption 3.3 and is positive definite. This time, for $\mathsf{X} \in \mathbb{R}^{N\times d}$, we define

$$\|\mathsf{X}\|_{W^{-1}-I} \coloneqq \sqrt{\operatorname{tr}(\mathsf{X}^\mathsf{T}(W^{-1}-I)\mathsf{X})}.$$

1. Show that $\|\mathsf{X}\|_{W^{-1}-I} = 0$ if and only if X has identical row vectors.
2. Consider the following optimization problem

$$\min_{\mathsf{X}\in\mathbb{R}^{N\times d}} F(\mathsf{X}) + \frac{1}{2\eta}\|\mathsf{X}\|_{W^{-1}-I}^2,$$

where where F is defined by (3.1), and we assume each f_i is convex. Show that X^* is an optimal solution to this problem if and only if

$$\mathsf{X}^* = W(\mathsf{X}^* - \eta\nabla F(\mathsf{X}^*)).$$ □

Convergence for Strongly Convex Functions

One of the technical difficulties of deriving convergence results of DGD for strongly convex functions is to establish the boundedness of $\|\Delta(t)\|_F$. When $f_i : \mathbb{R}^d \to \mathbb{R}$ is strongly convex,

we cannot assume that f_i is Lipschitz continuous over $\mathbb{R}^d$ anymore, as we have the following proposition:

Proposition 3.1. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex. Then there does not exist $G > 0$ such that f is G-Lipschitz continuous over $\mathbb{R}^d$.*

Exercise 3.8. Prove Proposition 3.1. Note that we do not assume the function f is differentiable. □

Fortunately, as we have seen, the aforementioned issue can be addressed when we employ a constant step size $\eta_t = \eta$ and adopt the perspective of viewing DGD as gradient descent for minimizing $F(\mathsf{X}) + \|\mathsf{X}\|_{I-W}^2/(2\eta)$. Specifically, if each f_i is μ-strongly convex and L-smooth, then Assumption 3.2 will be automatically satisfied. Therefore, assuming that each agent starts from the same initial point $x_i(0) = x_0$, we have

$$\|\Delta(t)\|_F^2 \le 2LND^2, \qquad D \coloneqq \sqrt{f(x_0) - \frac{1}{N}\sum_{i=1}^N f_i^\ell}.$$

Lemma 3.1 then gives

$$\|\mathsf{E}(t)\|_F \le \frac{\sqrt{2LN}D\eta}{1-\sigma}.$$

We then analyze the convergence of the averaged iterate $\bar{x}(t)$. Let x^* be the minimizer of $f(x)$, and let $\eta \in (0, 2/(\mu + L)]$. Denoting $\alpha = 1 - \eta\mu$, we have

$$\begin{aligned}
\|\bar{x}(t+1) - x^*\| &\le \|\bar{x}(t) - \eta\nabla f(\bar{x}(t)) - x^*\| + \eta\|\nabla f(\bar{x}(t)) - \bar{g}(t)\| \\
&\le \alpha\|\bar{x}(t) - x^*\| + \eta\frac{1}{N}\left\|\sum_{i=1}^N (\nabla f_i(\bar{x}(t)) - \nabla f_i(x_i(t)))\right\| \\
&\le \alpha\|\bar{x}(t) - x^*\| + \eta\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\bar{x}(t)) - \nabla f_i(x_i(t))\| \\
&\le \alpha\|\bar{x}(t) - x^*\| + \eta\frac{L}{N}\sum_{i=1}^N \|\bar{x}(t) - x_i(t)\| \\
&\le \alpha\|\bar{x}(t) - x^*\| + \eta\frac{L}{\sqrt{N}}\|\mathsf{E}(t)\|_F \le \alpha\|\bar{x}(t) - x^*\| + \frac{\eta^2 L\sqrt{2L}D}{1-\sigma},
\end{aligned}$$

where the second inequality follows from Theorem 1.4. Consequently, it can be shown by mathematical induction that

$$\|\bar{x}(t) - x^*\| \le (1-\eta\mu)^t\|x_0 - x^*\| + \frac{\eta L\sqrt{2L}D}{\mu(1-\sigma)}, \tag{3.14}$$

which establishes the convergence of DGD for strongly convex functions with constant step sizes.

It turns out that the issue of bounding $\|\Delta(t)\|_F$ can also be addressed when we employ diminishing step sizes. We refer interested readers to [Choi and Kim, 2022, Choi, 2023] for the analysis with diminishing step sizes.

3.5 A Brief Discussion on the Complexity and Scalability

We provide a brief discussion on the complexity and scalability of the DGD algorithm. For simplicity of analysis we only consider the constant step size setting.

The Convex and Smooth Case. Assuming that the conditions of Theorem 3.2 (or the first part of Corollary 3.1 are satisfied, we have the bound

$$\min_{\tau\in\{1,\ldots,t\}} f(\bar{x}(\tau)) - f(x^*) \leq \frac{C_1}{\eta t} + \frac{C_2(\eta L)^2}{(1-\sigma^2)^2},$$

where C_1 and C_2 are some positive quantities that we assume to be constant. It's not hard to see that, if we plan ahead the total number of iterations of DGD to be T, and choose the step size η to be

$$\eta = \frac{\alpha}{L}\cdot\frac{(1-\sigma^2)^{2/3}}{T^{1/3}}$$

for a numerical constant $\alpha \in (0, 1]$, then

$$\min_{\tau=1,\ldots,T} f(\bar{x}(\tau)) - f(x^*) \leq O\bigg(\frac{1}{((1-\sigma^2)T)^{2/3}}.\bigg)$$

Consequently, to achieve $\min_{\tau=1,\ldots,T} f(\bar{x}(\tau)) - f(x^*) \leq \epsilon$, the iteration and communication complexity of DGD can be upper bounded by

$$O\bigg(\frac{1}{(1-\sigma^2)\epsilon^{3/2}}\bigg).$$

The above complexity bound also provides information on the scalability of the DGD algorithm. Especially, since $\frac{1-\sigma^2}{\ln(1/\sigma)} \to 2$ as $\sigma \uparrow 1$, we see that the complexity bound of DGD has similar dependence on σ compared with the basic consensus method for distributed averaging.

The Strongly Convex and Smooth Case. Without loss of generality we may assume $\eta \leq 2/(\mu + L)$. In this case $\alpha = 1 - \eta\mu$, and by (3.14) and $1 - x \leq \mathrm{e}^{-x}$, we have

$$\|\bar{x}(t) - x^*\| \leq C_1\mathrm{e}^{-\eta\kappa Lt} + \frac{C_2\eta L}{\kappa(1-\sigma)},$$

where C_1 and C_2 are some positive quantities that we assume to be constant, and $\kappa = \mu/L$. Now, suppose we fix the total number of iterations to be a sufficiently large T, and choose the constant step size η as

$$\eta = \frac{\ln T}{\kappa LT}.$$

Then,

$$\|\bar{x}(T) - x^*\| \leq \frac{C_1}{T} + \frac{C_2\ln T}{\kappa^2(1-\sigma)T} = O\left(\frac{\ln T}{\kappa^2(1-\sigma)T}\right).$$

As a result, we can see that, in order to achieve $\|\bar{x}(t)-x^*\| \le \epsilon$, the iteration and communication complexity can be bounded by (see Exercise 3.9)

$$O\left(\frac{\ln(1/\epsilon)}{\kappa^2(1-\sigma)\epsilon}\right).$$

Since $\frac{1-\sigma}{\ln(1/\sigma)} \to 1$ as $\sigma \uparrow 1$, we see that the complexity bound of DGD has similar dependence on σ compared with the basic consensus method for distributed averaging.

Exercise 3.9. Let $p > 0$ be arbitrary, and let $f : [1, +\infty) \to \mathbb{R}$ be defined by

$$f(t) = \frac{\ln t}{t^p}.$$

1. Find t_0 such that $f(t)$ is strictly decreasing over $[t_0, +\infty)$.
2. Let $h : (0, f(t_0)] \to [t_0, +\infty)$ be the inverse function of f on $[t_0, +\infty)$. Show that

$$\lim_{\epsilon \downarrow 0} \frac{h(\epsilon)}{(\epsilon^{-1}\ln(1/\epsilon))^{1/p}}$$

exists, and find its value. □

3.6 Some Extensions

We now give brief introductions to two extensions of the decentralized gradient descent method.

Decentralized subgradient descent. The decentralized subgradient descent method applies to consensus optimization problems in which the cost functions are convex but possibly nonsmooth. The iterations of decentralized subgradient descent are given by

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

where $g_i(t)$ is an arbitrary element in the subdifferential of the local cost function $\partial f_i(x_i(t))$. It also admits diffusion variants just like DGD.

To establish convergence guarantees, we can assume each $f_i : \mathbb{R}^d \to \mathbb{R}$ to be convex and G-Lipschitz continuous; we also assume $\|\mathsf{E}(0)\|_F = 0$ for simplicity. Then $\|g_i(t)\| \le G$ for all i and t, and thus

$$\|\Delta(t)\|_F = \left\| \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right) \begin{bmatrix} g_1(t)^\mathsf{T} \\ \vdots \\ g_N(t)^\mathsf{T} \end{bmatrix} \right\| \le \left\| \begin{bmatrix} g_1(t)^\mathsf{T} \\ \vdots \\ g_N(t)^\mathsf{T} \end{bmatrix} \right\| \le \sqrt{N}G,$$

implying that Lemma 3.1 and the results of Exercise 3.4 can be applied. To study the convergence of $\bar{x}(t)$, by denoting $\bar{g}(t) = \frac{1}{N}\sum_{i=1}^{N} g_i(t)$, we can still have

$$\frac{1}{2}\|\bar{x}(t+1) - x^*\|^2 = \eta_t \langle \bar{g}(t), x^* - \bar{x}(t+1)\rangle + \frac{1}{2}\|\bar{x}(t) - x^*\|^2 - \frac{1}{2}\|\bar{x}(t+1) - \bar{x}(t)\|^2. \tag{3.8}$$

This time, to bound $\langle \bar{g}(t), x^* - \bar{x}(t+1)\rangle$, we note that

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

while for the quantity $\langle \bar{g}(t), x^* - \bar{x}(t)\rangle$, we have

$$\begin{aligned}
\langle \bar{g}(t), x^* - \bar{x}(t)\rangle &= \frac{1}{N}\sum_{i=1}^{N}\langle g_i(t), x^* - x_i(t)\rangle + \frac{1}{N}\sum_{i=1}^{N}\langle g_i(t), x_i(t) - \bar{x}(t)\rangle \\
&\le \frac{1}{N}\sum_{i=1}^{N}(f_i(x^*) - f_i(x_i(t))) + \frac{1}{N}\sum_{i=1}^{N}\|g_i(t)\|\|x_i(t) - \bar{x}(t)\| \\
&\le f(x^*) - f(\bar{x}(t)) + \frac{1}{N}\sum_{i=1}^{N}|f_i(\bar{x}(t)) - f_i(x_i(t))| + \frac{G}{N}\sum_{i=1}^{N}\|x_i(t) - \bar{x}(t)\| \\
&\le f(x^*) - f(\bar{x}(t)) + \frac{2G}{N}\sum_{i=1}^{N}\|x_i(t) - \bar{x}(t)\| \le f(x^*) - f(\bar{x}(t)) + \frac{2G}{\sqrt{N}}\|\mathsf{E}(t)\|_F.
\end{aligned}$$

Combining the above derivations, it can be shown that

$$\eta_t(f(\bar{x}(t)) - f(x^*)) \le \frac{1}{2}\left(\|\bar{x}(t) - x^*\|^2 - \|\bar{x}(t+1) - x^*\|^2\right) + \frac{\eta_t^2 G^2}{2} + \frac{2G}{\sqrt{N}}\eta_t\|\mathsf{E}(t)\|_F.$$

and by taking the telescoping sum and using the results of Exercise 3.4, we get

$$\frac{\sum_{\tau=0}^{t-1}\eta_\tau(f(\bar{x}(\tau)) - f(x^*))}{\sum_{\tau=0}^{t-1}\eta_\tau} \le \frac{\|\bar{x}(0) - x^*\|^2 + G^2\left(\frac{4}{1-\sigma} + 1\right)\sum_{\tau=0}^{t-1}\eta_\tau^2}{2\sum_{\tau=0}^{t-1}\eta_\tau}.$$

The following theorem summarizes the convergence results of the decentralized subgradient descent method.

Theorem 3.3. *Suppose each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and G-Lipschitz continuous, and suppose each agent starts from the same initial point $x_i(0) = x_0$.*

1. *Suppose a constant step size $\eta_t = \eta$ is chosen for the decentralized subgradient descent method. Then we have*

$$\frac{1}{t}\sum_{\tau=0}^{t-1}(f(\bar{x}(\tau)) - f(x^*)) \le \frac{\|x_0 - x^*\|^2}{\eta t} + \frac{5\eta G^2}{1-\sigma},$$

and

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

2. Suppose we choose diminishing step sizes $\eta_t \propto 1/\sqrt{t+1}$. Then

$$\min_{\tau=0,\ldots,t-1}(f(\bar{x}(\tau)) - f(x^*)) \leq O\left(\frac{\ln t}{\sqrt{t}}\right),$$

and

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

Decentralized stochastic gradient descent. Just as the name suggests, the decentralized stochastic gradient descent (DSGD) method applies to consensus optimization problems where only stochastic gradients of local cost functions can be accessed. The iterations of DSGD are given by

$$x_i(t+1) = \sum_{i=1}^{N} W_{ij}x_j(t) - \eta_t g_i(t),$$

where $g_i(t)$ in this time is a stochastic gradient of f_i that satisfies $\mathbb{E}[g_i(t)|x_i(t)] = \nabla f_i(x_i(t))$. DSGD also admits diffusion variants.

We leave convergence analysis of DSGD to the interested readers.

Notes on References

The decentralized gradient descent method can date back to the 1980s [Tsitsiklis et al., 1986]; this pioneering work proposed the fundamental framework of DGD and its variants, and also established preliminary almost sure convergence guarantees. One of the standard references on the decentralized gradient descent method is the seminal work [Nedić and Ozdaglar, 2009], but it considered possibly nonsmooth convex cost functions and time-varying communication networks. The Master's thesis [Chen, 2012] is also an important early work on DGD and its extensions. The diffusion method for consensus optimization was introduced in [Chen and Sayed, 2012], and the work [Sayed, 2014] provided a relatively comprehensive account of the diffusion method. The paper [Yuan et al., 2016] established convergence results of DGD with constant step sizes. The two recent works [Choi and Kim, 2022, Choi, 2023] established that DGD for strongly convex objectives with diminishing step sizes $\eta_t \propto \mu/(t+t_0)$ achieves convergence rate $\|\bar{x}(t) - x^*\| \leq O(1/t)$. Some related works on decentralized stochastic gradient descent include [Ram et al., 2010], [Jakovetic et al., 2018], [Yuan et al., 2019], etc.; see also the references therein.

Lemma 3.2 was adapted from [Qu and Li, 2018]. The materials of Section 3.4 are mostly based on the results in [Yuan et al., 2016].

Bibliography

[Chen, 2012] Chen, I.-A. (2012). Fast distributed first-order methods. Master's thesis, Massachusetts Institute of Technology.

[Chen and Sayed, 2012] Chen, J. and Sayed, A. H. (2012). Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305.

[Choi, 2023] Choi, W. (2023). A tight bound on the stepsize of the decentralized gradient descent. *arXiv preprint arXiv:2303.05755*.

[Choi and Kim, 2022] Choi, W. and Kim, J. (2022). On the convergence of decentralized gradient descent with diminishing stepsize, revisited. *arXiv preprint arXiv:2203.09079*.

[Jakovetic et al., 2018] Jakovetic, D., Bajovic, D., Sahu, A. K., and Kar, S. (2018). Convergence rates for distributed stochastic optimization over random networks. In *Proceedings of the 57th IEEE Conference on Decision and Control (CDC)*, pages 4238–4245.

[Jakovetic et al., 2011] Jakovetic, D., Xavier, J., and Moura, J. M. (2011). Fast distributed gradient methods. Available at `https://arxiv.org/abs/1112.2972v3`.

[Nedić and Ozdaglar, 2009] Nedić, A. and Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61.

[Qu and Li, 2018] Qu, G. and Li, N. (2018). Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260.

[Ram et al., 2010] Ram, S. S., Nedić, A., and Veeravalli, V. V. (2010). Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 147(3):516–545.

[Sayed, 2014] Sayed, A. H. (2014). Adaptation, learning, and optimization over networks. *Foundations and Trends® in Machine Learning*, 7(4-5):311–801.

[Tsitsiklis et al., 1986] Tsitsiklis, J., Bertsekas, D., and Athans, M. (1986). Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812.

[Yuan et al., 2019] Yuan, K., Alghunaim, S. A., Ying, B., and Sayed, A. H. (2019). On the performance of exact diffusion over adaptive networks. In *Proceedings of the 58th IEEE Conference on Decision and Control (CDC)*, pages 4898–4903.

[Yuan et al., 2016] Yuan, K., Ling, Q., and Yin, W. (2016). On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854.

Chapter 4

Gradient Tracking for Distributed Optimization

4.1 Motivation and the Algorithm

In the last chapter, we have presented and discussed the decentralized gradient descent method, one of the most fundamental algorithms in the area of distributed optimization. The DGD method is relatively straightforward to implement and has convergence guarantees. However, as we have seen, there are gaps between DGD and the centralized gradient descent in terms of the convergence rate guarantees. The most notable gap is in the convergence rate when the objective function is smooth (see Table 4.1). Also, the convergence of DGD also requires some additional conditions. Since about 2015, there have been multiple works proposing novel distributed optimization algorithms that attempt to address these issues, and the focus of this chapter is to introduce some of these distributed algorithms.

	Convex		Convex & smooth		Strongly convex & smooth	
	Rate	Complexity	Rate	Complexity	Rate	Complexity
Centralized (sub)GD	$O\left(\frac{\ln t}{\sqrt{t}}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\epsilon}\right)$	$O\big((1-c\kappa)^t\big)$	$O\left(\kappa \ln \frac{1}{\epsilon}\right)$
DGD	$O\left(\frac{\ln t}{\sqrt{t}}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$	$O\left(\frac{\ln t}{t^{2/3}}\right)$	$O\left(\frac{1}{\epsilon^{3/2}}\right)$	$O\left(\frac{1}{t^2}\right)$	$O\left(\frac{1}{\sqrt{\epsilon}}\right)$

Table 4.1: Comparison of the convergence rates and complexities of DGD and the centralized gradient descent method. The convergence rates and complexities are evaluated in terms of $\min_{0\leq\tau\leq t-1} f(\bar{x}(\tau)) - f(x^*)$.

In order to address the aforementioned issues, let's first see what intuitively causes the slow

convergence of DGD. We have seen that for DGD, the averaged iterate $\bar{x}(t)$ satisfies

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

and when all $x_i(t)$ are close to each other, i.e., approximate consensus has been reached, we have $\frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_i(t)) \approx \nabla f(\bar{x}(t))$, and therefore $\bar{x}(t)$ evolves just like centralized gradient descent with time-varying step sizes η_t. Recall that in Chapter 1, all convergence rate results for (deterministic) gradient descent were derived with a constant positive step size. Therefore, in order for DGD to achieve comparable convergence rates with centralized GD, the step size η_t need to have a positive lower bound throughout the iterations (see also Exercise 4.1). However, as we have seen, if the step size η_t has a positive lower bound, then $\mathsf{X}^* = \mathbf{1}x^{*\mathsf{T}}$ in general will not be a fixed point of the DGD iteration $\mathsf{X}(t+1) = W\mathsf{X}(t) - \eta_t \nabla F(\mathsf{X}(t))$. This dilemma forces us to either give up accurate convergence to the optimal solution, or to use diminishing step sizes that result in slower convergence.

Note the above observations also suggest a potential direction for addressing the issue of slow convergence of DGD: If we can design the distributed optimization iteration so that

i) $\bar{x}(t+1) - \bar{x}(t) - \eta \nabla f(\bar{x}(t))$ approaches zero sufficiently fast, where η is a *constant* step size; and

ii) $\mathsf{X}^* = \mathbf{1}x^{*\mathsf{T}}$ is a fixed point of the iteration,

then the resulting algorithm should be more likely to achieve comparable convergence rates with centralized gradient descent.

Exercise 4.1. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth, and let x^* be the minimizer of f over $\mathbb{R}^d$. Consider the (centralized) gradient descent iteration

$$x(t+1) = x(t) - \eta_t \nabla f(x(t)).$$

1. Show that
$$\|x(t+1) - x^*\| \geq (1 - \eta_t L)\|x(t) - x^*\|.$$

2. Suppose $\lim_{t\to\infty} \eta_t = 0$, and $x(0) \neq x^*$. Show that for any $\delta \in (0, 1)$, there exists $T \in \mathbb{N}$ such that for any $t \geq T$, $\|x(t+1) - x^*\| \geq \delta \|x(t) - x^*\|$. Consequently, the gradient descent method does not achieve linear convergence. □

In this chapter, we first introduce the *gradient tracking* algorithm for distributed optimization. The intuition behind the gradient tracking algorithm is to introduce an auxiliary local variable $g_i(t)$ that serves as a local estimate of the global gradient. Specifically, the gradient tracking

algorithm is given by[1]

$$x_i(t+1) = \sum_{i=1}^{N} W_{ij}\, x_j(t) - \eta\, g_i(t) \tag{4.1a}$$

$$g_i(t+1) = \sum_{i=1}^{N} W_{ij}\, g_j(t) + \nabla f_i(x_i(t+1)) - \nabla f_i(x_i(t)), \tag{4.1b}$$

with the initialization $g_i(0) = \nabla f_i(x_i(0))$. If we introduce the notations

$$\mathsf{X}(t) = \begin{bmatrix} — & x_1(t)^\mathsf{T} & — \\ & \vdots & \\ — & x_N(t)^\mathsf{T} & — \end{bmatrix}, \qquad \mathsf{G}(t) = \begin{bmatrix} — & g_1(t)^\mathsf{T} & — \\ & \vdots & \\ — & g_N(t)^\mathsf{T} & — \end{bmatrix},$$

then it's not hard to see that the gradient tracking iterations can be compactly written as

$$\begin{aligned} &\mathsf{X}(t+1) = W\mathsf{X}(t) - \eta\mathsf{G}(t), \\ &\mathsf{G}(t+1) = W\mathsf{G}(t) + \nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t)), \qquad \mathsf{G}(0) = \nabla F(\mathsf{X}(0)). \end{aligned} \tag{4.2}$$

Note that in the gradient tracking algorithm (4.1), each agent needs to exchange *two* vectors of dimension d with its neighbors for each iteration.

Exercise 4.2. Consider the gradient tracking algorithm (4.1).

1. Show that

$$\frac{1}{N}\sum_{i=1}^{N} g_i(t) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_i(t)) \quad \text{and} \quad \bar{x}(t+1) = \bar{x}(t) - \eta \sum_{i=1}^{N} \nabla f_i(x_i(t))$$

   for all $t \in \mathbb{N}$, where $\bar{x}(t) \coloneqq \frac{1}{N}\sum_{i=1}^{N} x_i(t)$.

2. Let x^* be a minimizer of f over $\mathbb{R}^d$. Find a fixed point of the gradient tracking iteration (4.2). □

We now provide some intuitive explanations of why $g_i(t)$ can serve as a local estimate of the global gradient $\nabla f(\bar{x}(t))$:

[1] This version in the lecture notes is adopted from [Qu and Li, 2018, Nedić et al., 2017]. The gradient tracking method also has "diffusion" variants, one of which is given by

$$\begin{aligned} x_i(t+1) &= \textstyle\sum_{j=1}^{N} W_{ij}(x_j(t) - \eta\, g_j(t)), \\ g_i(t+1) &= \textstyle\sum_{j=1}^{N} W_{ij}(g_j(t) + \nabla f_j(x_j(t+1)) - \nabla f_j(x_j(t))). \end{aligned}$$

See [Di Lorenzo and Scutari, 2016, Xu et al., 2015] for more details.

1. As Exercise 4.2 shows, we have $\frac{1}{N}\sum_{i=1}^N g_i(t) = \frac{1}{N}\sum_{i=1}^N \nabla f_i(x_i(t))$. Therefore, if all $x_i(t)$ and $g_i(t)$ have reached approximate consensus, then

$$g_i(t) \approx \frac{1}{N}\sum_{i=1}^N g_i(t) = \frac{1}{N}\sum_{i=1}^N \nabla f_i(x_i(t)) \approx \frac{1}{N}\sum_{i=1}^N \nabla f_i(\bar{x}(t)) = \nabla f(\bar{x}(t)),$$

   i.e., $g_i(t)$ approximates the global gradient $\nabla f(\bar{x}(t))$.

2. The above derivation requires that $x_i(t)$ and $g_i(t)$ have reached approximate consensus. From (4.1), it can be seen that, when $\bar{x}(t)$ approaches x^*, as long as $g_i(t)$ approximates the true gradient $\nabla f(\bar{x}(t))$, the increment $-\eta g_i(t)$ in (4.1a) will be very small, which then has little effects on the consensus procedure on $x_i(t)$ in (4.1a). Additionally, if all $x_i(t)$ have also reached approximate consensus, $x_i(t+1) - x_i(t)$ will be small, and then the smoothness of f_i implies that $\nabla f_i(x_i(t+1)) - \nabla f_i(x_i(t))$ will be small, meaning that the increment $\nabla f_i(x_i(t+1)) - \nabla f_i(x_i(t))$ in (4.1b) will have little effects on the consensus procedure on $g_i(t)$. As a result, both $x_i(t)$ and $g_i(t)$ will reach consensus.

The above explanations seem to fall into circular reasoning. Nevertheless, we shall see in our theoretical analysis how the above intuitions can be justified rigorously.

4.2 Convergence Analysis: The Smooth and Strongly Convex Case

For analysis purposes, we introduce the notations

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

We also use $\mathsf{E}_x(t)$ and $\mathsf{E}_g(t)$ to denote the consensus errors:

$$\mathsf{E}_x(t) \coloneqq \begin{bmatrix} (x_1(t) - \bar{x}(t))^\mathsf{T} \\ \vdots \\ (x_N(t) - \bar{x}(t))^\mathsf{T} \end{bmatrix} = \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)\mathsf{X}(t),$$

$$\mathsf{E}_g(t) \coloneqq \begin{bmatrix} (g_1(t) - \bar{g}(t))^\mathsf{T} \\ \vdots \\ (g_N(t) - \bar{g}(t))^\mathsf{T} \end{bmatrix} = \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)\mathsf{G}(t).$$

As Exercise 4.2 shows, we have

$$\bar{g}(t) = \frac{1}{N}\sum_{i=1}^N \nabla f_i(x_i(t)),$$

and

$$\bar{x}(t+1) = \bar{x}(t) - \eta\bar{g}(t) = \bar{x}(t) - \eta\frac{1}{N}\sum_{i=1}^N \nabla f_i(x_i(t)).$$

For analysis purposes, we impose the condition $\eta L \leq 1$ on the constant step size η. We shall see that a stronger condition will be imposed on η for establishing convergence.

We first derive a bound on $\|\bar{g}(t) - \nabla f(\bar{x}(t))\|$ that will be useful in subsequent analysis.

$$\begin{aligned}\|\bar{g}(t) - \nabla f(\bar{x}(t))\| &= \left\| \frac{1}{N} \sum_{i=1}^{N} (\nabla f_i(x_i(t)) - \nabla f_i(\bar{x}(t))) \right\| \\ &\leq \frac{1}{N} \sum_{i=1}^{N} \|\nabla f_i(x_i(t)) - \nabla f_i(\bar{x}(t))\| \\ &\leq \frac{L}{N} \sum_{i=1}^{N} \|x_i(t) - \bar{x}(t)\| \leq \frac{L}{\sqrt{N}} \|\mathsf{E}_x(t)\|_F.\end{aligned} \tag{4.3}$$

To begin our analysis, we first consider the evolution of the consensus errors $\|\mathsf{E}_x(t)\|_F$ and $\|\mathsf{E}_g(t)\|_F$. We have

$$\begin{aligned}\|\mathsf{E}_x(t+1)\|_F &= \left\| \left(I - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right) (W\mathsf{X}(t) - \eta \mathsf{G}(t)) \right\|_F \\ &= \left\| \left(W - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \mathsf{X}(t) - \eta \left(I - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \mathsf{G}(t) \right\|_F \\ &= \left\| \left(W - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \mathsf{E}_x(t) - \eta \mathsf{E}_g(t) \right\|_F \\ &\leq \left\| \left(W - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \mathsf{E}_x(t) \right\|_F + \eta \left\| \mathsf{E}_g(t) \right\|_F \\ &\leq \sigma \|\mathsf{E}_x(t)\|_F + \eta \|\mathsf{E}_g(t)\|_F.\end{aligned}$$

While for $\mathsf{E}_g(t)$, we have

$$\begin{aligned}\|\mathsf{E}_g(t+1)\|_F &= \left\| \left(I - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \left(W\mathsf{G}(t) + \nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t)) \right) \right\|_F \\ &= \left\| \left(W - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \mathsf{E}_g(t) + \left(I - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \left(\nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t)) \right) \right\|_F \\ &\leq \left\| \left(W - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \mathsf{E}_g(t) \right\|_F + \left\| \left(I - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \left(\nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t)) \right) \right\|_F \\ &\leq \sigma \|\mathsf{E}_g(t)\|_F + \left\| \nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t)) \right\|_F,\end{aligned}$$

where we used $\left\| I - \frac{1}{N} \mathbf{1}\mathbf{1}^\mathsf{T} \right\|_2 = 1$. Note that

$$\begin{aligned}&\left\| \nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t)) \right\|_F \\ &\leq L \|\mathsf{X}(t+1) - \mathsf{X}(t)\|_F = L \left\| (W - I)\mathsf{X}(t) - \eta \mathsf{G}(t) \right\|_F \\ &= L \left\| (W - I)\mathsf{E}_x(t) - \eta \mathsf{G}(t) \right\|_F \leq L \left\| (W - I)\mathsf{E}_x(t) \right\|_F + \eta L \|\mathsf{G}(t)\|_F \\ &\leq 2L \|\mathsf{E}_x(t)\|_F + \eta L \|\mathsf{G}(t)\|_F,\end{aligned}$$

and to bound $\|\mathsf{G}(t)\|_F$, we have

$$\begin{aligned}\|\mathsf{G}(t)\|_F &= \|1 \cdot \nabla f(\bar{x}(t))^\mathsf{T} + \mathbf{1}(\bar{g}(t) - \nabla f(\bar{x}(t))^\mathsf{T} + \mathsf{E}_g(t)\|_F \\ &\leq \|1 \cdot \nabla f(\bar{x}(t))^\mathsf{T}\|_F + \left\|\mathbf{1}(\bar{g}(t) - \nabla f(\bar{x}(t))^\mathsf{T}\right\|_F + \|\mathsf{E}_g(t)\|_F \\ &= \sqrt{N}\|\nabla f(\bar{x}(t)) - \nabla f(x^*)\| + \sqrt{N}\,\|\bar{g}(t) - \nabla f(\bar{x}(t))\| + \|\mathsf{E}_g(t)\|_F \\ &\leq \sqrt{N}L\|\bar{x}(t) - x^*\| + L\|\mathsf{E}_x(t)\|_F + \|\mathsf{E}_g(t)\|_F.\end{aligned}$$

Therefore,

$$\begin{aligned}\|\mathsf{E}_g(t+1)\|_F &\leq (\sigma + \eta L)\|\mathsf{E}_g(t)\|_F + (2 + \eta L)L\|\mathsf{E}_x(t)\|_F + \eta L^2\sqrt{N}\|\bar{x}(t) - x^*\| \\ &\leq (\sigma + \eta L)\|\mathsf{E}_g(t)\|_F + 3L\|\mathsf{E}_x(t)\|_F + \eta L^2\sqrt{N}\|\bar{x}(t) - x^*\|\end{aligned}$$

Finally, regarding the quantity $\|\bar{x}(t) - x^*\|$, we have

$$\begin{aligned}\|\bar{x}(t+1) - x^*\| = \|\bar{x}(t) - \eta\bar{g}(t) - x^*\| &\leq \|\bar{x}(t) - \eta\nabla f(\bar{x}(t)) - x^*\| + \|\eta\nabla f(\bar{x}(t)) - \eta\bar{g}(t)\| \\ &\leq (1 - \eta\mu)\|\bar{x}(t) - x^*\| + \frac{\eta L}{\sqrt{N}}\|\mathsf{E}_x(t)\|_F,\end{aligned}$$

where in the first inequality we used Theorem 1.4 and the fact that $|\eta L - 1| \leq 1 - \eta\mu$ for $\eta \leq 1/L$.

Summarizing the previous results, we obtain the following inequalities:

1. $\|\mathsf{E}_x(t+1)\|_F \leq \sigma\|\mathsf{E}_x(t)\|_F + \eta\|\mathsf{E}_g(t)\|_F$.
2. $\|\mathsf{E}_g(t+1)\|_F \leq (\sigma + \eta L)\|\mathsf{E}_g(t)\|_F + 3L\|\mathsf{E}_x(t)\|_F + \eta L^2\sqrt{N}\|\bar{x}(t) - x^*\|$.
3. $\|\bar{x}(t+1) - x^*\| \leq (1 - \eta\mu)\|\bar{x}(t) - x^*\| + \frac{\eta L}{\sqrt{N}}\|\mathsf{E}_x(t)\|_F$.

In fact, these inequalities can be regarded as rigorous and precise formulations of the intuitive explanations given in the last section.

We may rewrite the three inequalities in the following compact form:

$$\begin{bmatrix} L^{-1}\|\mathsf{E}_g(t+1)\|_F \\ \|\mathsf{E}_x(t+1)\|_F \\ \sqrt{N}\|\bar{x}(t+1) - x^*\| \end{bmatrix} \leq \begin{bmatrix} \sigma + \eta L & 3 & \eta L \\ \eta L & \sigma & 0 \\ 0 & \eta L & 1 - \eta\mu \end{bmatrix} \begin{bmatrix} L^{-1}\|\mathsf{E}_g(t)\|_F \\ \|\mathsf{E}_x(t)\|_F \\ \sqrt{N}\|\bar{x}(t) - x^*\| \end{bmatrix}, \tag{4.4}$$

where the inequality is componentwise. Denote

$$M(\epsilon) \coloneqq \begin{bmatrix} \sigma + \epsilon & 3 & \epsilon \\ \epsilon & \sigma & 0 \\ 0 & \epsilon & 1 - \kappa\epsilon \end{bmatrix}.$$

Since all relevant entries in (4.4) are nonnegative, we can apply mathematical induction and get

$$\begin{bmatrix} L^{-1}\|\mathsf{E}_g(t)\|_F \\ \|\mathsf{E}_x(t)\|_F \\ \sqrt{N}\|\bar{x}(t) - x^*\| \end{bmatrix} \leq M(\eta L)^t \begin{bmatrix} L^{-1}\|\mathsf{E}_g(0)\|_F \\ \|\mathsf{E}_x(0)\|_F \\ \sqrt{N}\|\bar{x}(0) - x^*\| \end{bmatrix}. \tag{4.5}$$

The following lemmas help us handle the power of $M(\eta L)$ by noticing that it is a nonnegative matrix satisfying the conditions of the Perron-Frobenius Theorem 2.2.

Lemma 4.1. *For any $\epsilon \in (0,1]$, there exist $u(\epsilon), v(\epsilon) \in \mathbb{R}^3$ such that*

$$\lim_{t\to\infty} \left(\frac{M(\epsilon)}{\rho(M(\epsilon))}\right)^t = \frac{u(\epsilon)v(\epsilon)^\mathsf{T}}{u(\epsilon)^\mathsf{T}v(\epsilon)}. \tag{4.6}$$

Proof. It's not hard to check that for any $\epsilon \in (0,1]$, the matrix $M(\epsilon)$ satisfies the conditions of the Perron-Frobenius Theorem 2.2. Thus $\rho(M(\epsilon))$ is a simple eigenvalue of $M(\epsilon)$, and all other eigenvalues of $M(\epsilon)$ has magnitudes strictly less than $\rho(M(\epsilon))$. Consequently, by letting $u(\epsilon)$ and $v(\epsilon)$ be right and left eigenvectors of $M(\epsilon)$ associated with the eigenvalue $\rho(M(\epsilon))$ respectively, we can get (4.6); see [Horn and Johnson, 2013, Theorem 8.6.1], or the results derived in Exercise 2.4 and Lemma 2.1. □

It now becomes evident that we need to bound the spectral radius of $M(\epsilon)$ for $\epsilon \in (0,1]$.

Lemma 4.2. *Suppose $\epsilon \le 1$. Then*

$$\rho(M(\epsilon)) \le \max\left\{1 - \frac{\kappa\epsilon}{2}, \sigma + 5\sqrt{\frac{\epsilon}{\kappa}}\right\}.$$

Particularly, when $\epsilon \le \kappa\left(\frac{1-\sigma}{6}\right)^2$, we have $\rho(M(\epsilon)) = 1 - \frac{\kappa\epsilon}{2} < 1$.

Proof. First of all, we note that the matrix $M(\epsilon)$ has nonnegative entries and satisfies the condition of the Perron-Frobenius Theorem 2.2 when $\eta \le 1$. Thus we can conclude that $\rho(M(\epsilon))$ is an eigenvalue of $M(\epsilon)$, which further implies that it is the largest real root of the characteristic polynomial of $M(\epsilon)$, which we temporarily denote by $p(\lambda)$. Since the degree of $p(\lambda)$ is 3, as long as we find some $\overline{\lambda} \in \mathbb{R}$ such that $p(\overline{\lambda}) \ge 0$ and $p(\lambda)$ is increasing over $\lambda \in (\overline{\lambda}, +\infty)$, we can then conclude that $\overline{\lambda}$ is an upper bound of the largest real root of $p(\lambda)$, and thus is an upper bound of $\rho(M(\epsilon))$.

The characteristic polynomial $p(\lambda)$ is given by

$$\begin{aligned}
p(\lambda) = \det(\lambda I - M(\epsilon)) &= \begin{vmatrix} \lambda-\sigma-\epsilon & -3 & -\epsilon \\ -\epsilon & \lambda-\sigma & 0 \\ 0 & -\epsilon & \lambda-(1-\kappa\epsilon) \end{vmatrix} \\
&= (\lambda-\sigma-\epsilon)(\lambda-\sigma)(\lambda-(1-\kappa\epsilon)) - 3\epsilon(\lambda-(1-\kappa\epsilon)) - \epsilon^3 \\
&= [(\lambda-\sigma-\epsilon)(\lambda-\sigma) - 3\epsilon](\lambda-(1-\kappa\epsilon)) - \epsilon^3 \\
&= (\lambda-\lambda_+)(\lambda-\lambda_-)(\lambda-(1-\kappa\epsilon)) - \epsilon^3,
\end{aligned}$$

where $\lambda_\pm$ are the two real roots of the equation $(\lambda-\sigma-\epsilon)(\lambda-\sigma) - 3\epsilon = 0$ and are given by

$$\lambda_\pm = \sigma + \frac{\epsilon}{2} \pm \sqrt{3\epsilon + \frac{\epsilon^2}{4}}.$$

We can see that both of $\lambda_\pm$ are bounded above by

$$\lambda_\pm \le \sigma + \sqrt{\epsilon}\left(\frac{\sqrt{\epsilon}}{2} + \sqrt{3 + \frac{\epsilon}{4}}\right) \le \sigma + \sqrt{\epsilon}\left(\frac{1}{2} + \sqrt{3 + \frac{1}{4}}\right) < \sigma + 3\sqrt{\epsilon}.$$

Therefore, $p(\lambda)$ is increasing when $\lambda > \max\{1-\kappa\epsilon, \sigma+3\sqrt{\epsilon}\} \geq \max\{1-\kappa\epsilon, \lambda_+, \lambda_-\}$. All we need to do now is to find $\overline{\lambda} \geq \max\{1-\kappa\epsilon, \sigma+3\sqrt{\epsilon}\}$ such that $p(\overline{\lambda}) \geq 0$.

We now claim that the choice

$$\overline{\lambda} = \max\left\{1-\frac{\kappa\epsilon}{2}, \sigma+5\sqrt{\frac{\epsilon}{\kappa}}\right\}$$

does the work. Indeed, since both of $\lambda_\pm$ are less than $\sigma+3\sqrt{\epsilon}$, we have

$$\begin{aligned}
p(\overline{\lambda}) &= (\overline{\lambda}-\lambda_+)(\overline{\lambda}-\lambda_-)(\overline{\lambda}-(1-\kappa\epsilon)) - \epsilon^3 \\
&\geq (\overline{\lambda}-\sigma-3\sqrt{\epsilon})^2(\overline{\lambda}-(1-\kappa\epsilon)) - \epsilon^3 \\
&\geq \left(5\sqrt{\frac{\epsilon}{\kappa}}-3\sqrt{\epsilon}\right)^2\left(1-\frac{\kappa\epsilon}{2}-(1-\kappa\epsilon)\right) - \epsilon^3 \\
&= \left[\left(\frac{5}{\sqrt{\kappa}}-3\right)^2\frac{\kappa}{2}-\epsilon\right]\epsilon^2 \geq \left[\left(\frac{5-3}{\sqrt{\kappa}}\right)^2\frac{\kappa}{2}-\epsilon\right]\epsilon^2 = (2-\epsilon)\epsilon^2 > 0.
\end{aligned}$$

Our proof is now complete. □

Combining the previous results, we see that, if we choose the step size η to satisfy

$$\eta \leq \frac{\kappa}{L}\left(\frac{1-\sigma}{6}\right)^2 = \frac{\mu}{L^2}\left(\frac{1-\sigma}{6}\right)^2,$$

then $\rho(M(\eta L)) \leq 1-\frac{\eta\mu}{2} < 1$, and from (4.5) we can get

$$\frac{1}{\rho(M(\eta L))^t}\begin{bmatrix} L^{-1}\|\mathsf{E}_g(t)\|_F \\ \|\mathsf{E}_x(t)\|_F \\ \sqrt{N}\|\bar{x}(t)-x^*\| \end{bmatrix} \leq \left(\frac{M(\eta L)}{\rho(M(\eta L))}\right)^t \begin{bmatrix} L^{-1}\|\mathsf{E}_g(0)\|_F \\ \|\mathsf{E}_x(0)\|_F \\ \sqrt{N}\|\bar{x}(0)-x^*\| \end{bmatrix}$$

By taking the norm and letting $t \to \infty$, the right-hand side becomes

$$\left\| \frac{u(\eta L)v(\eta L)^\mathsf{T}}{u(\eta L)^\mathsf{T}v(\eta L)} \begin{bmatrix} L^{-1}\|\mathsf{E}_g(0)\|_F \\ \|\mathsf{E}_x(0)\|_F \\ \sqrt{N}\|\bar{x}(0)-x^*\| \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} L^{-1}\|\mathsf{E}_g(0)\|_F \\ \|\mathsf{E}_x(0)\|_F \\ \sqrt{N}\|\bar{x}(0)-x^*\| \end{bmatrix} \right\|.$$

Therefore we can conclude that

$$\left\| \begin{bmatrix} L^{-1}\|\mathsf{E}_g(t)\|_F \\ \|\mathsf{E}_x(t)\|_F \\ \sqrt{N}\|\bar{x}(t)-x^*\| \end{bmatrix} \right\| \leq O(\rho(M(\eta L))^t) \leq O\left(\left(1-\frac{\eta\mu}{2}\right)^t\right).$$

We summarize the convergence results of the gradient tracking algorithm 4.1 in the following theorem.

Theorem 4.1. *Suppose each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth, and the global objective function f is μ-strongly convex. Let the step size satisfy*

$$\eta \leq \alpha\,\frac{\mu}{L^2}\left(\frac{1-\sigma}{6}\right)^2,$$

for some $\alpha \in (0,1]$, and let $x_i(t)$, $g_i(t)$ be generated by the gradient tracking algorithm 4.1. Then $\|\bar{x}(t) - x^*\|$ and the consensus errors $\frac{1}{N}\sum_{i=1}^N \|x_i(t) - \bar{x}(t)\|^2$, $\frac{1}{N}\sum_{i=1}^N \|g_i(t) - \nabla f(\bar{x}(t))\|^2$ all achieve the linear convergence rate

$$O\left(\left[1 - \frac{\alpha}{2}\left(\frac{\kappa(1-\sigma)}{6}\right)^2\right]^t\right).$$

4.3 Convergence Analysis: The Smooth and Convex Case

We now turn to the analysis of the gradient tracking algorithm 4.1 for smooth and convex objective functions.

Recall that for the gradient tracking algorithm, we have

$$\bar{x}(t+1) = \bar{x}(t) - \eta\bar{g}(t) = \bar{x}(t) - \eta\,\frac{1}{N}\sum_{i=1}^N \nabla f_i(x_i(t)).$$

Therefore, Lemma 3.2 in Chapter 3 can be applied here, and from Lemma 3.2, we see that the critical step is to derive an upper bound on the quantity $\sum_{\tau=0}^{t-1}\|\mathsf{E}_x(\tau)\|_F^2$. We therefore start from the inequality

$$\|\mathsf{E}_x(t+1)\|_F \le \sigma\|\mathsf{E}_x(t)\|_F + \eta\|\mathsf{E}_g(t)\|_F$$

which has been derived in the last section. By taking the square and using the AM–GM inequality, we get

$$\begin{aligned}
\|\mathsf{E}_x(t+1)\|_F^2 \le\ & \left(1 + \frac{(1-\sigma)(1+2\sigma)}{2\sigma^2}\right)\sigma^2\|\mathsf{E}_x(t)\|_F^2 + \left(1 + \frac{2\sigma^2}{(1-\sigma)(1+2\sigma)}\right)\eta^2\|\mathsf{E}_g(t)\|_F^2 \\
=\ & \frac{1+\sigma}{2}\|\mathsf{E}_x(t)\|_F^2 + \eta^2\,\frac{1+\sigma}{(1-\sigma)(1+2\sigma)}\|\mathsf{E}_g(t)\|_F^2 \\
\le\ & \frac{1+\sigma}{2}\|\mathsf{E}_x(t)\|_F^2 + \eta^2\,\frac{1}{1-\sigma}\|\mathsf{E}_g(t)\|_F^2,
\end{aligned}$$

where we used $(1+\sigma)/(1+2\sigma) \le 1$ in the last step. Then for $\mathsf{E}_g(t)$, note that in the last section we have shown

$$\begin{aligned}
\|\mathsf{E}_g(t+1)\|_F &\le \sigma\|\mathsf{E}_g(t)\|_F + \|\nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t))\|_F \\
&\le \sigma\|\mathsf{E}_g(t)\|_F + 2L\|\mathsf{E}_x(t)\|_F + \eta L\|\mathsf{G}(t)\|_F.
\end{aligned}$$

This time we bound $\|\mathsf{G}(t)\|_F$ differently:

$$\begin{aligned}
\|\mathsf{G}(t)\|_F =\ & \left\|\mathbf{1}\bar{g}(t)^\mathsf{T} + \mathsf{E}_g(t)\right\|_F \\
\le\ & \|\mathbf{1}\bar{g}(t)\|_F + \|\mathsf{E}_g(t)\|_F = \sqrt{N}\|\bar{g}(t)\| + \|\mathsf{E}_g(t)\|_F.
\end{aligned}$$

As a result, if we impose the condition $\eta L \le (1-\sigma)/3$ on the step size η, we get

$$\begin{aligned}
\|\mathsf{E}_g(t+1)\|_F \le\ & (\sigma + \eta L)\,\|\mathsf{E}_g(t)\|_F + 2L\|\mathsf{E}_x(t)\|_F + \eta L\sqrt{N}\|\bar{g}(t)\| \\
\le\ & \frac{1+2\sigma}{3}\|\mathsf{E}_g(t)\|_F + 2L\|\mathsf{E}_x(t)\|_F + \eta L\sqrt{N}\|\bar{g}(t)\|.
\end{aligned}$$

By taking the square and using the AM–GM inequality, we have

$$\begin{aligned}\|\mathsf{E}_g(t+1)\|_F^2 \le{}& \left(1+\frac{(1-\sigma)(7+8\sigma)}{2(1+2\sigma)^2}\right)\left(\frac{1+2\sigma}{3}\right)^2\|\mathsf{E}_g(t)\|_F^2 \\ &+\left(1+\frac{2(1+2\sigma)^2}{(1-\sigma)(7+8\sigma)}\right)\left(2L\|\mathsf{E}_x(t)\|_F+\eta L\sqrt{N}\|\bar{g}(t)\|\right)^2 \\ ={}& \frac{1+\sigma}{2}\|\mathsf{E}_g(t)\|_F^2+\frac{9(1+\sigma)}{(1-\sigma)(7+8\sigma)}\left(2L\|\mathsf{E}_x(t)\|_F+\eta L\sqrt{N}\|\bar{g}(t)\|\right)^2 \\ \le{}& \frac{1+\sigma}{2}\|\mathsf{E}_g(t)\|_F^2+\frac{3}{2(1-\sigma)}(8L^2\|\mathsf{E}_x(t)\|_F^2+2\eta^2L^2N\|\bar{g}(t)\|^2),\end{aligned}$$

where we used $9(1+\sigma)/(7+8\sigma) < 3/2$ for $\sigma \in [0,1]$. Summarizing the previous results, we have the following two inequalities:

$$\begin{aligned}\|\mathsf{E}_x(t+1)\|_F^2 &\le \frac{1+\sigma}{2}\|\mathsf{E}_x(t)\|_F^2+\frac{\eta^2}{1-\sigma}\|\mathsf{E}_g(t)\|_F^2, \\ \|\mathsf{E}_g(t+1)\|_F^2 &\le \frac{1+\sigma}{2}\|\mathsf{E}_g(t)\|_F^2+\frac{12L^2}{1-\sigma}\|\mathsf{E}_x(t)\|_F^2+\frac{3\eta^2L^2N}{1-\sigma}\|\bar{g}(t)\|^2.\end{aligned}$$

Similarly, these two inequalities can be written in the following compact form:

$$\begin{bmatrix}\|\mathsf{E}_x(t+1)\|_F^2 \\ \frac{\eta}{2\sqrt{3}L}\|\mathsf{E}_g(t+1)\|_F^2\end{bmatrix} \le P(\eta L)\begin{bmatrix}\|\mathsf{E}_x(t)\|_F^2 \\ \frac{\eta}{2\sqrt{3}L}\|\mathsf{E}_g(t)\|_F^2\end{bmatrix}+\frac{\sqrt{3}\eta^3LN}{2(1-\sigma)}\begin{bmatrix}0 \\ \|\bar{g}(t)\|^2\end{bmatrix}, \tag{4.7}$$

where the inequality is componentwise, and we denote

$$P(\epsilon) \coloneqq \begin{bmatrix}\frac{1+\sigma}{2} & \frac{2\sqrt{3}\epsilon}{1-\sigma} \\ \frac{2\sqrt{3}\epsilon}{1-\sigma} & \frac{1+\sigma}{2}\end{bmatrix}.$$

Here we deliberately scale $\|\mathsf{E}_g(t)\|_F^2$ so that the matrix $P(\eta L)$ becomes real symmetric, which simplifies the analysis of the spectrum of $P(\epsilon)$ as well as the asymptotic behavior of $P(\epsilon)^t$ as $t \to \infty$. Indeed, we have the following lemma:

Lemma 4.3. *We have*

$$\|P(\epsilon)\|_2 \le \frac{2+\sigma}{3}$$

whenever $0 < \epsilon \le ((1-\sigma)/5)^2$.

Proof. The two eigenvalues of $P(\epsilon)$ are explicitly given by

$$\lambda_\pm = \frac{1+\sigma}{2} \pm \frac{2\sqrt{3}\epsilon}{1-\sigma}.$$

Then since $P(\epsilon)$ is real symmetric, the desired bound on $\|P(\epsilon)\|_2$ then directly follows from $\|P(\epsilon)\|_2 \le \lambda_+$ and $0 < \epsilon \le ((1-\sigma)/5)^2$. □

We can now plug the bound derived in Lemma 4.3 into (4.7) to get

$$\left\|\begin{bmatrix}\|\mathsf{E}_x(t+1)\|_F^2 \\ \frac{\eta}{2\sqrt{3}L}\|\mathsf{E}_g(t+1)\|_F^2\end{bmatrix}\right\| \le \frac{2+\sigma}{3}\left\|\begin{bmatrix}\|\mathsf{E}_x(t)\|_F^2 \\ \frac{\eta}{2\sqrt{3}L}\|\mathsf{E}_g(t)\|_F^2\end{bmatrix}\right\|+\frac{\sqrt{3}\eta^3LN}{2(1-\sigma)}\|\bar{g}(t)\|^2,$$

assuming that $\eta L \leq ((1-\sigma)/5)^2$. By using mathematical induction, we then obtain

$$\left\| \begin{bmatrix} \|\mathsf{E}_x(t)\|_F^2 \\ \frac{\eta}{2\sqrt{3}L}\|\mathsf{E}_g(t)\|_F^2 \end{bmatrix} \right\| \leq \left(\frac{2+\sigma}{3}\right)^t NE_0^2 + \frac{\sqrt{3}\eta^3 LN}{2(1-\sigma)} \sum_{\tau=0}^{t-1} \left(\frac{2+\sigma}{3}\right)^{t-1-\tau} \|\bar{g}(\tau)\|^2,$$

where we denote

$$E_0 \coloneqq \sqrt{\frac{1}{N}\left(\|\mathsf{E}_x(0)\|_F^2 + \frac{\eta}{2\sqrt{3}L}\|\mathsf{E}_g(0)\|_F^2\right)},$$

and used the fact that the ℓ_2-norm of a vector in $\mathbb{R}^n$ is less than or equal to its ℓ_1-norm. By summing over t, we have

$$\sum_{\tau=0}^{t-1} \|\mathsf{E}_x(\tau)\|_F^2 \leq NE_0^2 \sum_{\tau=0}^{t-1} \left(\frac{2+\sigma}{3}\right)^{\tau} + \frac{\sqrt{3}\eta^3 LN}{2(1-\sigma)} \sum_{\tau=0}^{t-1}\sum_{s=0}^{\tau-1} \left(\frac{2+\sigma}{3}\right)^{\tau-1-s} \|\bar{g}(s)\|^2.$$

Note that the second term on the right-hand side can be bounded by interchanging the two summations:

$$\begin{aligned} \sum_{\tau=0}^{t-1}\sum_{s=0}^{\tau-1} \left(\frac{2+\sigma}{3}\right)^{\tau-1-s} \|\bar{g}(s)\|^2 &= \sum_{s=0}^{t-2} \|\bar{g}(s)\|^2 \sum_{\tau=s+1}^{t-1} \left(\frac{2+\sigma}{3}\right)^{\tau-1-s} \\ &\leq \sum_{s=0}^{t-2} \|\bar{g}(s)\|^2 \sum_{\tau=s+1}^{\infty} \left(\frac{2+\sigma}{3}\right)^{\tau-1-s} \leq \frac{3}{1-\sigma} \sum_{\tau=0}^{t-1} \|\bar{g}(\tau)\|^2. \end{aligned}$$

Therefore

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

Recall that when the step size is constant, Lemma 3.2 gives

$$\frac{1}{t}\sum_{\tau=1}^{t}(f(\bar{x}(\tau)) - f(x^*)) \leq \frac{\|\bar{x}(0)-x^*\|^2}{2\eta t} + \frac{L}{2Nt}\sum_{\tau=0}^{t-1}\|\mathsf{E}(\tau)\|_F^2 + \frac{(\eta L-1)\sum_{\tau=0}^{t-1}\|\bar{x}(\tau+1)-\bar{x}(\tau)\|^2}{2\eta t}.$$

By plugging in the bound on $\sum_{\tau=0}^{t-1}\|\mathsf{E}(\tau)\|_F^2$ and noting that $\|\bar{x}(\tau+1)-\bar{x}(\tau)\|^2 = \eta^2\|\bar{g}(\tau)\|^2$, we get

$$\begin{aligned} \frac{1}{t}\sum_{\tau=1}^{t}(f(\bar{x}(\tau)) - f(x^*)) &\leq \frac{1}{t}\left[\frac{\|\bar{x}(0)-x^*\|^2}{2\eta} + \frac{3LE_0^2}{2(1-\sigma)} + \frac{\eta}{2}\left(\frac{3\sqrt{3}(\eta L)^2}{2(1-\sigma)^2} + \eta L - 1\right)\sum_{\tau=0}^{t-1}\|\bar{g}(\tau)\|^2\right] \\ &\leq \frac{1}{t}\left[\frac{\|\bar{x}(0)-x^*\|^2}{2\eta} + \frac{3LE_0^2}{2(1-\sigma)}\right], \end{aligned} \tag{4.8}$$

where the last step follows because $\eta L \leq ((1-\sigma)/5)^2$ implies $\frac{3\sqrt{3}(\eta L)^2}{2(1-\sigma)^2} + \eta L - 1 \leq 0$.

After deriving the convergence rate for $f(\bar{x}(t)) - f(x^*)$, we proceed to bound the consensus errors, which we leave as a series of exercises for the readers.

Exercise 4.3. Let $(c_t)_{t\in\mathbb{N}}$ be a sequence of nonnegative real numbers that satisfies $\sum_{t=0}^{\infty} c_t < +\infty$. Show that

$$\lim_{t\to\infty}\left(t\cdot\min_{0\le\tau\le t-1} c_\tau\right)=0.$$

In other words, we have

$$\min_{0\le\tau\le t-1} c_\tau = o\left(\frac{1}{t}\right)$$

as $t\to\infty$. □

Exercise 4.4. 1. Suppose η satisfies $\eta L \le ((1-\sigma)/5)^2$. Show that

$$\frac{3\sqrt{3}(\eta L)^2}{2(1-\sigma)^2}+\eta L-1<-\frac{4}{5}.$$

This is a very loose bound, but suffices for subsequent derivations.

2. Show that

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

(Hint: Use (4.8) and note that $f(x)-f(x^*)\ge 0$ for any $x\in\mathbb{R}^d$.)

3. Use the above result to derive a bound on $\sum_{\tau=0}^{t-1}\|\mathsf{E}_x(\tau)\|_F^2$.

4. Show that

$$\min_{\tau=0,\dots,t-1}\|\mathsf{E}_x(\tau)\|_F^2\le o\left(\frac{1}{t}\right).$$

□

We can now summarize the convergence results of gradient tracking for smooth and convex objective functions in the following theorem:

Theorem 4.2. *Suppose each $f_i:\mathbb{R}^d\to\mathbb{R}$ is convex and L-smooth, and suppose $x^*\in\mathbb{R}^d$ is a minimizer of f over $\mathbb{R}^d$. Let $x_i(t)$, $t\ge 0$ be generated by the gradient tracking algorithm 4.1 with step size η satisfying*

$$\eta=\frac{\alpha}{L}\left(\frac{1-\sigma}{5}\right)^2$$

for $\alpha\in(0,1]$. Then

$$\frac{1}{t}\sum_{\tau=1}^{t}(f(\bar{x}(\tau))-f(x^*))\le\frac{1}{t}\left[\frac{25L\|\bar{x}(0)-x^*\|^2}{2\alpha(1-\sigma)^2}+\frac{3LE_0^2}{2(1-\sigma)}\right],$$

and

$$\min_{\tau=0,\dots,t-1}\frac{1}{N}\sum_{i=1}^{N}\|x_i(\tau)-\bar{x}(\tau)\|^2\le o\left(\frac{1}{t}\right),$$

where $\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$, *and*

$$E_0 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\|x_i(0) - \bar{x}(0)\|^2 + \frac{\eta}{2\sqrt{3}L}\left\|\nabla f_i(x_i(0)) - \frac{1}{N}\sum_{j=1}^{N}\nabla f_j(x_j(0))\right\|^2\right)}.$$

Summary of the Convergence and Complexity Results

The main convergence rate results derived above can be summarized as follows:

- Convex & smooth:
$$\frac{1}{t}\sum_{\tau=1}^{t}(f(\bar{x}(\tau)) - f(x^*)) \leq O\left(\frac{1}{(1-\sigma)^2 t}\right).$$
- Strongly convex & smooth:
$$f(\bar{x}(t)) - f(x^*) \leq O\left(\left[1 - c\kappa^2(1-\sigma)^2\right]^t\right),$$
where $c > 0$ is a numerical constant, and $\kappa = \mu/L$ is the reciprocal condition number.

From these convergence rate results, we can further derive the iteration complexities of the gradient tracking algorithm (4.1):

- Convex & smooth: The number of iterations T to achieve $\frac{1}{T}\sum_{\tau=1}^{T}(f(\bar{x}(\tau)) - f(x^*)) \leq \epsilon$ is bounded by
$$O\left(\frac{1}{(1-\sigma)^2\epsilon}.\right)$$
- Strongly convex & smooth: The number of iterations T to achieve $f(\bar{x}(T)) - f(x^*)) \leq \epsilon$ is bounded by
$$O\left(\frac{1}{\kappa^2(1-\sigma)^2}\ln\frac{1}{\epsilon}\right).$$

We can see that, the convergence rates as well as the iteration complexity bounds now match those of the centralized gradient descent method, in terms of the dependencies on t and ϵ, respectively. On the other hand, there is still room for improvement:

1. The iteration complexities of the gradient tracking algorithm (4.1) now scale with the network's topology as $O((1-\sigma)^{-2})$. Compared to DGD and the consensus method for distributed averaging, the scalability of the gradient tracking algorithm (4.1) seems to be worse.
2. For strongly convex and smooth objective functions, our analysis shows that the iteration complexity of the gradient tracking algorithm (4.1) scales with the condition number as $O(\kappa^{-2})$, which is worse than the centralized gradient descent method $O(\kappa^{-1})$.
3. The gradient tracking algorithm (4.1) is based on the vanilla gradient descent method. It is expected that, by incorporating Nesterov's acceleration in the design of our distributed optimization algorithm, we can achieve faster convergence and lower complexities.

4.4 Other Gradient-Tracking-Type Distributed Optimization Algorithms

In this section, we present some other distributed optimization algorithms that employ constant step sizes and achieve theoretical convergence rates comparable to their centralized counterpart.

- EXTRA: EXTRA (short for exact first-order algorithm) [Shi et al., 2015a] is perhaps the first distributed optimization algorithm that bridges the gap in the convergence rates compared to the centralized counterpart. The iterations of EXTRA are given by

$$x_i(t+2) = x_i(t+1) + \sum_{j=1}^{N} W_{ij}x_j(t+1) - \sum_{j=1}^{N} \tilde{W}_{ij}x_j(t) - \eta(\nabla f_i(x_i(t+1)) - \nabla f_i(x_i(t))),$$

  or, written compactly,

$$\mathsf{X}(t+2) = (I+W)\mathsf{X}(t+1) - \tilde{W}\mathsf{X}(t) - \eta(\nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t))).$$

  Here $W, \tilde{W} \in \mathbb{R}^{N\times N}$ are two real symmetric weight matrices that satisfy

  1. $\text{null}(I-\tilde{W}) \supseteq \text{null}(W-\tilde{W}) = \text{span}\{\mathbf{1}\}$, and
  2. $\tilde{W} \succ 0$ and $W \preceq \tilde{W} \preceq (I+W)/2$.

  A common choice for $\tilde{W}$ is $\tilde{W} = (I+W)/2$, and in this case the above two conditions reduce to $-I \prec W \preceq I$ and $\text{null}(I-W) = \text{span}\{\mathbf{1}\}$, which will be satisfied if W is a Metropolis weight matrix.

- Exact diffusion/NIDS: The paper [Yuan et al., 2019a] proposes the exact diffusion algorithm, while the paper [Li et al., 2019] proposed the NIDS (short for network independent step-size) algorithm. The two algorithms coincide when the communication graph is undirected and the local cost functions are all smooth:

$$x_i(t+2) = \sum_{j=1}^{N} \overline{W}_{ij}(2x_j(t+1) - x_j(t) - \eta(\nabla f_j(x_j(t+1)) - \nabla f_j(x_j(t))),)$$

  or

$$\mathsf{X}(t+2) = \overline{W}(2\mathsf{X}(t+1) - \mathsf{X}(t) - \eta(\nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t))),$$

  where $\overline{W} = (I+W)/2$ and W is a real symmetric weight matrix satisfying $W\mathbf{1} = \mathbf{1}$ and $\left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right\|_2 < 1$.

  Exact diffusion/NIDS can be regarded as a diffusion variant of EXTRA with $\tilde{W} = (I+W)/2$. In addition, both exact diffusion/NIDS and EXTRA only require one round of communication between agents per iteration.

The following theorem provides the convergence results of exact diffusion/NIDS for strongly convex and smooth local objectives:

Theorem 4.3 ([Xu et al., 2021]). *Let $W \in \mathbb{R}^{N\times N}$ be a real symmetric weight matrix satisfying $W\mathbf{1} = \mathbf{1}$ and $\sigma = \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right\|_2 < 1$. Suppose each local cost function f_i is μ-strongly convex*

and L-smooth. Then by choosing the step size to be

$$\eta = \frac{2}{L+\mu},$$

the exact diffusion/NIDS algorithm achieves the convergence rate

$$\frac{1}{N}\sum_{i=1}^{N}\|x_i(t)-x^*\|^2 \le O\left(\left[\max\left\{\left(\frac{L-\mu}{L+\mu}\right)^2, \frac{1+\sigma}{2}\right\}\right]^t\right).$$

As a corollary, the iteration complexity of exact diffusion/NIDS for strongly convex and smooth local cost functions is given by

$$O\left(\max\left\{\frac{1}{\kappa}, \frac{1}{1-\sigma}\right\}\ln\frac{1}{\epsilon}\right),$$

where $\kappa = \mu/L$. The convergence rate of exact diffusion/NIDS for convex and smooth local cost functions can also be found in [Xu et al., 2021], which we omit here.

A Primal-Dual Perspective of Gradient-Tracking-Type Algorithms

In Section 3.4, we have seen that DGD with constant step sizes can be viewed as doing gradient descent for minimizing certain cost function. In this subsection, we present a similar perspective for grading-tracking-type algorithms.

Assume throughout this subsection that $W \in \mathbb{R}^{N\times N}$ is real symmetric and positive semidefinite, and satisfies $\sigma := \|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\|_2 < 1$. Recall that the optimization problem for the analysis of DGD is

$$\min_{\mathsf{X}\in\mathbb{R}^{N\times d}} F(\mathsf{X}) + \frac{1}{2\eta}\|\mathsf{X}\|_{I-W}^2,$$

where $F(\mathsf{X}) = \sum_{i=1}^N f_i(x_i)$ with x_i^T being the i'th column of X, and $\|\mathsf{X}\|_{I-W}^2 = \mathrm{tr}(\mathsf{X}^\mathsf{T}(I-W)\mathsf{X})$. We see that $\frac{1}{2\eta}\|\mathsf{X}\|_{I-W}^2$ can be viewed as a smooth penalty function that drives X to consensus. However, this penalty function does not result in optimal solutions that have strictly reached consensus. One commonly used approach to remedy this issue is to consider the *augmented Lagrangian*

$$\mathscr{L}_\alpha(\mathsf{X},\mathsf{Y}) = F(\mathsf{X}) + \frac{1}{\alpha}\langle \mathsf{Y}, U\mathsf{X}\rangle + \frac{1}{2\alpha}\|\mathsf{X}\|_V^2,$$

which is the Lagrangian function corresponding to the following constrained optimization problem:

$$\begin{aligned}\min_{\mathsf{X}\in\mathbb{R}^d}\ & F(\mathsf{X}) + \frac{1}{2\alpha}\|\mathsf{X}\|_V^2\\ \text{s.t.}\ & \alpha^{-1}U\mathsf{X} = 0.\end{aligned}$$

Here U and V are matrices satisfying $\mathrm{null}\, U = \mathrm{null}\, V = \mathrm{span}\{\mathbf{1}\}$, V is positive semidefinite, and $\|X\|_V := \sqrt{\mathrm{tr}(\mathsf{X}^\mathsf{T}V\mathsf{X})}$ is the seminorm associated with V. We then apply the primal-dual gradient method (or one of its variants) to find a saddle point of the augmented Lagrangian $\mathscr{L}_\alpha(\mathsf{X},\mathsf{Y})$, in the hope that the resulting algorithm can be implemented in a distributed fashion, and will converge to an optimal solution with good performance.

Exercise 4.5. Suppose $A \in \mathbb{R}^{N\times N}$ is real symmetric and satisfies $0 \preceq A \preceq I$. Let

$$A = \sum_{i=1}^{N} \lambda_i u_i u_i^\mathsf{T}$$

be the eigenvalue decomposition of A, where $\{u_i : i = 1, \dots, n\}$ forms an orthonormal basis of $\mathbb{R}^N$. Given any function $f : [0, 1] \to \mathbb{R}$, we define

$$f(A) = \sum_{i=1}^{N} f(\lambda_i) u_i u_i^\mathsf{T}.$$

1. Suppose $f(x) = 0$ if and only if $x = 0$. Prove that $\operatorname{null} f(A) = \operatorname{null} A$.
2. Suppose $f(x) > 0$ for all $x \in [0, 1]$. Prove that $f(A)$ is positive definite.
3. Let $W \in \mathbb{R}^{N\times N}$ be a real symmetric matrix satisfying $W\mathbf{1} = \mathbf{1}$ and $\left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 < 1$. Show that
$$\operatorname{null}(I - W) = \operatorname{null}\sqrt{I - W} = \operatorname{null}(I - W^2) = \operatorname{span}\{\mathbf{1}\}. \qquad \square$$

At first glance, the above primal-dual perspective does not seem to be quite related to the gradient-tracking algorithms we have presented. So next we show how we can derive those gradient-tracking-type algorithms from this perspective.

1. Let $U = I - W$ and $V = I - W^2$. We then have
$$\begin{aligned}
\nabla_{\mathsf{X}} \mathscr{L}_\alpha(\mathsf{X}, \mathsf{Y}) &= \nabla F(\mathsf{X}) + \frac{1}{\alpha}(I - W)\mathsf{Y} + \frac{1}{\alpha}(I - W^2)\mathsf{X}, \\
\nabla_{\mathsf{Y}} \mathscr{L}_\alpha(\mathsf{X}, \mathsf{Y}) &= \frac{1}{\alpha}(I - W)\mathsf{X}.
\end{aligned}$$
Therefore, if we apply the Gauss-Seidel variant of the primal-dual gradient method to the augmented Lagrangian
$$\begin{aligned}
\mathsf{X}(t+1) &= \mathsf{X}(t) - \eta \nabla_{\mathsf{X}} \mathscr{L}_\alpha(\mathsf{X}(t), \mathsf{Y}(t)) \\
\mathsf{Y}(t+1) &= \mathsf{Y}(t) + \eta \nabla_{\mathsf{Y}} \mathscr{L}_\alpha(\mathsf{X}(t+1), \mathsf{Y}(t))
\end{aligned} \tag{4.9}$$
with step size $\eta = \alpha$, we get the iterations
$$\begin{aligned}
\mathsf{X}(t+1) &= W^2\mathsf{X}(t) - \eta \nabla F(\mathsf{X}(t)) - (I - W)\mathsf{Y}(t), \\
\mathsf{Y}(t+1) &= \mathsf{Y}(t) + (I - W)\mathsf{X}(t+1).
\end{aligned}$$
The above iterations do not appear familiar to us at first glance. However, we notice that the second equality implies $W\mathsf{X}(t+1) + Y(t+1) = \mathsf{Y}(t) + \mathsf{X}(t+1)$, and by plugging the first equality into the right-hand side, we get
$$W\mathsf{X}(t+1) + \mathsf{Y}(t+1) = W(W\mathsf{X}(t) + \mathsf{Y}(t)) - \eta \nabla F(\mathsf{X}(t)).$$

As a result, if we initialize $\mathsf{Y}(0)$ by $\mathsf{Y}(0) = -W\mathsf{X}(0)$, we have

$$W\mathsf{X}(t) + \mathsf{Y}(t) = -\eta \sum_{\tau=0}^{t-1} W^{t-1-\tau} \nabla F(\mathsf{X}(\tau)),$$

and by multiplying the above equality with $I - W$, we get

$$\begin{aligned}
&(I-W)(W\mathsf{X}(t) + \mathsf{Y}(t)) \\
&= \eta \sum_{\tau=0}^{t-1} W^{t-\tau} \nabla F(\mathsf{X}(\tau)) - \eta \sum_{\tau=0}^{t-1} W^{t-1-\tau} \nabla F(\mathsf{X}(\tau)) \\
&= \eta \sum_{\tau=0}^{t-1} W^{t-\tau} \nabla F(\mathsf{X}(\tau)) - \eta \sum_{\tau=1}^{t} W^{t-\tau} \nabla F(\mathsf{X}(\tau-1)) \\
&= \eta \left[-\nabla F(\mathsf{X}(t-1)) + \sum_{\tau=1}^{t-1} W^{t-\tau} (\nabla F(\mathsf{X}(\tau)) - \nabla F(\mathsf{X}(\tau-1))) + W^t \nabla F(\mathsf{X}(0)) \right].
\end{aligned}$$

Therefore, if we define

$$\mathsf{G}(t) = \nabla F(\mathsf{X}(t)) + \frac{1}{\eta}(I-W)(W\mathsf{X}(t) + \mathsf{Y}(t)),$$

then we will get

$$\mathsf{G}(t) = \sum_{\tau=1}^{t} W^{t-\tau} (\nabla F(\mathsf{X}(\tau)) - \nabla F(\mathsf{X}(\tau-1)) + W^t \nabla F(\mathsf{X}(0)),$$

from which we can see that $\mathsf{G}(t)$ is just generated by the iteration (4.1b) with $\mathsf{G}(0) = \nabla F(\mathsf{X}(0))$. Moreover,

$$\begin{aligned}
\mathsf{X}(t+1) &= W^2 \mathsf{X}(t) - \eta \left(\mathsf{G}(t) - \frac{1}{\eta}(I-W)(W\mathsf{X}(t) + \mathsf{Y}(t)) \right) - (I-W)\mathsf{Y}(t), \\
&= W\mathsf{X}(t) - \eta \mathsf{G}(t),
\end{aligned}$$

which coincides with (4.1a). We see that we have just recovered the gradient tracking algorithm (4.1).

Exercise 4.6. Consider the following augmented Lagrangian

$$\mathscr{L}_\alpha(\mathsf{X}, \mathsf{Y}) = F(\mathsf{X}) + \frac{1}{\alpha} \langle \mathsf{Y}, (I-W)\mathsf{X} \rangle + \frac{1}{2\alpha} \|\mathsf{X}\|^2_{2(I-W)}.$$

Show that, if we apply the vanilla primal-dual gradient method

$$\begin{aligned}
\mathsf{X}(t+1) &= \mathsf{X}(t) - \eta \nabla_{\mathsf{X}} \mathscr{L}_\alpha(\mathsf{X}(t), \mathsf{Y}(t)) \\
\mathsf{Y}(t+1) &= \mathsf{Y}(t) + \eta \nabla_{\mathsf{Y}} \mathscr{L}_\alpha(\mathsf{X}(t), \mathsf{Y}(t))
\end{aligned}$$

with appropriate initialization, then we can also recover the gradient tracking algorithm (4.1). □

2. Let $U = \sqrt{\tilde{W} - W}$ and $V = I - \tilde{W}$, where W and $\tilde{W}$ are two real symmetric weight matrices satisfying the conditions of EXTRA. We then have

$$\nabla_{\mathsf{X}} \mathscr{L}_\alpha(\mathsf{X}, \mathsf{Y}) = \nabla F(\mathsf{X}) + \frac{1}{\alpha}\sqrt{\tilde{W} - W}\mathsf{Y} + \frac{1}{\alpha}(I - \tilde{W})\mathsf{X},$$
$$\nabla_{\mathsf{Y}} \mathscr{L}_\alpha(\mathsf{X}, \mathsf{Y}) = \frac{1}{\alpha}\sqrt{\tilde{W} - W}\mathsf{X}.$$

By applying the Gauss-Seidel variant of the primal-dual gradient method (4.9) with step size $\eta = \alpha$, we get

$$\mathsf{X}(t+1) = \tilde{W}\mathsf{X}(t) - \eta\nabla F(\mathsf{X}(t)) - \sqrt{\tilde{W} - W}\mathsf{Y}(t)$$
$$\mathsf{Y}(t+1) = \mathsf{Y}(t) + \sqrt{\tilde{W} - W}\mathsf{X}(t+1).$$

We then try to eliminate the variable $\mathsf{Y}(t)$. Note that

$$\begin{aligned}\mathsf{X}(t+2) &= \tilde{W}\mathsf{X}(t+1) - \eta\nabla F(\mathsf{X}(t+1)) - \sqrt{\tilde{W} - W}\mathsf{Y}(t+1)\\ &= W\mathsf{X}(t+1) - \eta\nabla F(\mathsf{X}(t+1)) - \sqrt{\tilde{W} - W}\mathsf{Y}(t).\end{aligned}$$

By subtracting the updating rule of $\mathsf{X}(t+1)$ from the above equality, we get

$$\mathsf{X}(t+2) = (I + W)\mathsf{X}(t+1) - \tilde{W}\mathsf{X}(t) - \eta(\nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t))),$$

which is just EXTRA.

> *Exercise* 4.7. Suppose $\overline{W} \in \mathbb{R}^{N\times N}$ is a real symmetric weight matrix satisfying
>
> i) $\overline{W}\mathbf{1} = \mathbf{1}$ and $\left\|\overline{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right\|_2 < 1$;
>
> ii) $\overline{W} \succ 0$.
>
> Show that, the weight matrices $\tilde{W} = \overline{W}^2$ and $W = 2\overline{W} - I$ satisfy the conditions of EXTRA. Furthermore, if we set $\mathsf{X}(1) = \mathsf{X}(0) - \eta\nabla F(\mathsf{X}(0))$ in the corresponding EXTRA, then the sequence $(\mathsf{X}(t))_{t\geq 0}$ generated by EXTRA coincides with the one generated by the gradient tracking algorithm (4.1) with $\overline{W}$ as the weight matrix. □

3. Let $U = \sqrt{(I - W)/2}$ and $V = (I - W)/2$, and this time we consider decomposing the augmented Lagrangian as follows:

$$\mathscr{L}_\alpha(\mathsf{X}, \mathsf{Y}) = F(\mathsf{X}) + \underbrace{\frac{1}{\alpha}\left\langle \mathsf{Y}, \sqrt{\frac{I - W}{2}}\mathsf{X}\right\rangle + \frac{1}{2\alpha}\|\mathsf{X}\|^2_{(I-W)/2}}_{R_\alpha(\mathsf{X},\mathsf{Y})}.$$

We then apply the following variant of primal-dual gradient method, where we set $\eta = \alpha$:

$$\begin{aligned}\mathsf{X}(t+1/2) &= \mathsf{X}(t) - \eta\nabla F(\mathsf{X}(t)),\\ \mathsf{X}(t+1) &= \mathsf{X}(t+1/2) - \eta\nabla_{\mathsf{X}} R_\eta(\mathsf{X}(t+1/2), \mathsf{Y}(t)),\\ \mathsf{Y}(t+1) &= \mathsf{Y}(t) + \eta\nabla\mathscr{L}_\eta(\mathsf{X}(t+1), \mathsf{Y}(t)).\end{aligned}$$

We obtain

$$\mathsf{X}(t+1) = \frac{I+W}{2}(\mathsf{X}(t) - \eta\nabla F(\mathsf{X}(t))) - \sqrt{\frac{I-W}{2}}\mathsf{Y}(t),$$
$$\mathsf{Y}(t+1) = \mathsf{Y}(t) + \sqrt{\frac{I-W}{2}}\mathsf{X}(t+1).$$

Similarly, we can eliminate the variable $\mathsf{Y}(t)$ and get

$$\mathsf{X}(t+2) = \frac{I+W}{2}(2\mathsf{X}(t+1) - \mathsf{X}(t) - \eta(\nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t)))),$$

which is just the exact diffusion/NIDS algorithm.

We can see that, the primal-dual perspective reveals certain underlying connections between gradient-tracking-type algorithms. This allows us to develop unified frameworks for analyzing the convergence of gradient-tracking-type algorithms. For example, the paper [Xu et al., 2021] proposes the following general algorithm:

$$\mathsf{X}(t+1) = A\mathsf{X}(t) - \eta B\nabla F(\mathsf{X}(t)) - \mathsf{Y}(t),$$
$$\mathsf{Y}(t) = \mathsf{Y}(t) + C\mathsf{X}(t+1),$$

which include gradient tracking (with symmetric weight matrices), EXTRA and exact diffusion/NIDS as its special cases. The paper then proves convergence results for the above general algorithm using operator splitting techniques, which includes the following result as a special case:

Theorem 4.4. *Suppose the weight matrix W is real symmetric, positive definite and satisfies $W\mathbf{1} = \mathbf{1}$ and $\sigma = \left\|W - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right\|_2 < 1$. Suppose each f_i is μ-strongly convex and L-smooth. Then with a properly chosen step size, the number of iterations needed for the gradient tracking algorithm (4.1) to achieve $\|\bar{x}(t) - x^*\|^2 \leq \epsilon$ can be upper bounded by*

$$O\left(\max\left\{\frac{1}{\kappa\lambda_{\min}(W^2)}, \frac{1}{(1-\sigma)^2}\right\}\ln\frac{1}{\epsilon}\right),$$

where $\kappa = \mu/L$.

We can see that, if we choose the weight matrix W to be $W = (I + \overline{W})/2$ with $\overline{W}$ being the lazy Metropolis weight matrix, then $\lambda_{\min}(W^2) \geq 1/4$, and the above complexity bound will be (asymptotically) better than the one $O\left(\frac{1}{\kappa^2(1-\sigma)^2}\right)$ implied by Theorem 4.1. On the other hand, we point out that the conditions imposed in Theorem 4.4 is stronger than the conditions in Theorem 4.1.

Notes on References

Our materials on the gradient tracking algorithm (4.1) were based on [Qu and Li, 2018] and [Nedić et al., 2017]; [Nedić et al., 2017] named the algorithm (4.1) DIGing (a distributed

inexact gradient method and a gradient tracking technique). The paper [Qu and Li, 2018] analyzed (4.1) for both convex and strongly convex cases but assumed fixed undirected communication graphs, while [Nedić et al., 2017] only analyzed the strongly convex case but considered time-varying communication graphs. The analysis of (4.1) for convex and smooth functions also employs techniques from [Tang et al., 2021].

We note that the distributed tracking technique employed by (4.1) appeared in the earlier work [Zhu and Martínez, 2010]. Diffusion variants of (4.1) appeared in [Xu et al., 2015] and [Di Lorenzo and Scutari, 2016]; the former paper also considered uncoordinated step sizes, while the latter one mainly focused on nonconvex problems.

EXTRA was proposed by [Shi et al., 2015a] and, as we have mentioned, is perhaps the first work on distributed optimization algorithms that successfully bridge the gap in the convergence rate compared to the centralized counterparts. EXTRA was generalized to PG-EXTRA for distributed composite optimization in [Shi et al., 2015b]. Exact diffusion was proposed and studied in [Yuan et al., 2019a, Yuan et al., 2019b], which also considered locally balanced weight matrices. NIDS was proposed in [Li et al., 2019], which studied the more general composite optimization setting. The paper [Xu et al., 2021] proposed a unified framework for gradient-tracking-type algorithms and provide convergence analysis leveraging the theory of operator splitting. For more details on the theory of operator splitting and its relation with distributed optimization, we refer to the book [Ryu and Yin, 2022].

We mention that the gradient-tracking-type algorithms presented in this chapter also have close relationship with the following continuous-time distributed algorithm [Wang and Elia, 2010, Gharesifard and Cortés, 2014]

$$\vdots \begin{array}{l} \dot{\mathsf{X}}(t) = -L\mathsf{X}(t) - L\mathsf{Y}(t) - \nabla F(\mathsf{X}(t)), \\ \dot{\mathsf{Y}}(t) = L\mathsf{X}(t), \end{array} \tag{4.10}$$

where $L \in \mathbb{R}^{N\times N}$ is the Laplacian matrix of the communication graph. (4.10) can be viewed as the primal-dual gradient flow of the augmented Lagrangian

$$\mathscr{L}(\mathsf{X},\mathsf{Y}) = F(\mathsf{X}) + \langle \mathsf{Y}, L\mathsf{X}\rangle + \frac{1}{2}\|\mathsf{X}\|_L^2.$$

On the other hand, (4.10) can also be regarded as incorporating the PI control strategy to the gradient flow $\dot{\mathsf{X}}(t) = -\nabla F(\mathsf{X}(t))$, where $L\mathsf{X}(t)$ is the proportional term and $L\mathsf{Y}(t)$ is the integral term, so that $\mathsf{X}(t)$ will be driven to consensus.

Bibliography

[Di Lorenzo and Scutari, 2016] Di Lorenzo, P. and Scutari, G. (2016). Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136.

[Gharesifard and Cortés, 2014] Gharesifard, B. and Cortés, J. (2014). Distributed continuous-time convex optimization on weight-balanced digraphs. *IEEE Transactions on Automatic Control*, 59(3):781–786.

[Horn and Johnson, 2013] Horn, R. A. and Johnson, C. R. (2013). *Matrix Analysis*. Cambridge University Press, 2nd edition.

[Li et al., 2019] Li, Z., Shi, W., and Yan, M. (2019). A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67(17):4494–4506.

[Nedić et al., 2017] Nedić, A., Olshevsky, A., and Shi, W. (2017). Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633.

[Qu and Li, 2018] Qu, G. and Li, N. (2018). Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260.

[Ryu and Yin, 2022] Ryu, E. K. and Yin, W. (2022). *Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators*. Cambridge University Press, Cambridge, UK.

[Shi et al., 2015a] Shi, W., Ling, Q., Wu, G., and Yin, W. (2015a). EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966.

[Shi et al., 2015b] Shi, W., Ling, Q., Wu, G., and Yin, W. (2015b). A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing*, 63(22):6013–6023.

[Tang et al., 2021] Tang, Y., Zhang, J., and Li, N. (2021). Distributed zero-order algorithms for nonconvex multiagent optimization. *IEEE Transactions on Control of Network Systems*, 8(1):269–281.

[Wang and Elia, 2010] Wang, J. and Elia, N. (2010). Control approach to distributed optimization. In *Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 557–561.

[Xu et al., 2021] Xu, J., Tian, Y., Sun, Y., and Scutari, G. (2021). Distributed algorithms for composite optimization: Unified framework and convergence analysis. *IEEE Transactions on Signal Processing*, 69:3555–3570.

[Xu et al., 2015] Xu, J., Zhu, S., Soh, Y. C., and Xie, L. (2015). Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *Proceedings of the 54th IEEE Conference on Decision and Control (CDC)*, pages 2055–2060.

[Yuan et al., 2019a] Yuan, K., Ying, B., Zhao, X., and Sayed, A. H. (2019a). Exact diffusion for distributed optimization and learning — Part I: Algorithm development. *IEEE Transactions on Signal Processing*, 67(3):708–723.

[Yuan et al., 2019b] Yuan, K., Ying, B., Zhao, X., and Sayed, A. H. (2019b). Exact diffusion for distributed optimization and learning — Part II: Convergence analysis. *IEEE Transactions on Signal Processing*, 67(3):724–739.

[Zhu and Martínez, 2010] Zhu, M. and Martínez, S. (2010). Discrete-time dynamic average consensus. *Automatica*, 46(2):322–329.

Chapter 5

Alternating Direction Method of Multipliers

5.1 Introduction to ADMM

The alternating direction method of multipliers (ADMM) is a popular device that can be employed for designing algorithms for large-scale or distributed convex optimization problems. In this section, we provide an introduction to ADMM with some simple examples.

We start with a relatively simple problem setup. Consider the following convex optimization problem:

$$\begin{aligned} \min_{x_1\in\mathbb{R}^{d_1}, x_2\in\mathbb{R}^{d_2}} \quad & f_1(x_1)+f_2(x_2) \\ \text{s.t.} \quad & A_1x_1+A_2x_2=b, \end{aligned} \tag{5.1}$$

where $f_i : \mathbb{R}^{d_i} \to \mathbb{R}\cup\{+\infty\}$, $i=1,2$ are convex functions, A_1 and A_2 are real matrices of appropriate dimensions, and b is a real vector. We define the *augmented Lagrangian* of this optimization problem by

$$\mathscr{L}_\rho(x_1,x_2,y)=f_1(x_1)+f_2(x_2)+\langle y, A_1x_1+A_2x_2-b\rangle+\frac{\rho}{2}\|A_1x_1+A_2x_2-b\|^2.$$

We can see that, compared with the ordinary Lagrangian function, we add a penalty term $\frac{\rho}{2}\|A_1x_1+A_2x_2-b\|^2$ in the augmented Lagrangian, where $\rho>0$ is a positive parameter. The iterations of ADMM are given by

$$\begin{aligned} x_1(t+1) &\in \arg\min_{x_1}\mathscr{L}_\rho(x_1,x_2(t),y(t)), \\ x_2(t+1) &\in \arg\min_{x_2}\mathscr{L}_\rho(x_1(t+1),x_2,y(t)), \\ y(t+1) &= y(t)+\eta(A_1x_1+A_2x_2-b), \end{aligned}$$

or, equivalently,

$$\begin{aligned} x_1(t+1) &\in \operatorname*{arg\,min}_{x_1} \left(f_1(x_1) + \langle A_1^{\mathsf{T}} y(t), x_1 \rangle + \frac{\rho}{2} \|A_1 x_1 + A_2 x_2(t) - b\|^2 \right) \\ x_2(t+1) &\in \operatorname*{arg\,min}_{x_2} \left(f_2(x_2) + \langle A_2^{\mathsf{T}} y(t), x_2 \rangle + \frac{\rho}{2} \|A_1 x_1(t+1) + A_2 x_2 - b\|^2 \right) \\ y(t+1) &= y(t) + \eta \left(A_1 x_1(t+1) + A_2 x_2(t+1) - b \right). \end{aligned}$$

ADMM is closely related to the *dual ascent* method and the *method of multipliers*.

Dual Ascent, Method of Multipliers, and ADMM

Consider the equality-constrained convex optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f(x) \\ \text{s.t.} \quad & Ax = b, \end{aligned} \tag{5.2}$$

where $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a convex function, $A \in \mathbb{R}^{p \times d}$ and $b \in \mathbb{R}^p$. We always assume that the epigraph of f is nonempty and closed. The dual problem for (5.2) is given by

$$\max_{y \in \mathbb{R}^p} \quad g(y), \qquad g(y) = \inf_{x \in \mathbb{R}^d} \left(f(x) + \langle y, Ax - b \rangle \right).$$

It's not hard to show that g is a concave function. We also introduce the notion of *convex conjugate* of the convex function f, defined by

$$f^*(z) \coloneqq \sup_{x \in \mathbb{R}^d} \left(\langle z, x \rangle - f(x) \right).$$

It's not hard to see that f^* is a convex function that can take values in $\mathbb{R} \cup \{+\infty\}$, and the dual objective function g can be written as

$$g(y) = -f^*(-A^{\mathsf{T}} y) - \langle b, y \rangle.$$

The *dual ascent* method is based on the following observation, which can be derived as a corollary of [Rockafellar, 1970, Theorems 23.5 and 23.9]:

Lemma 5.1. *Let g denote the objective function of the dual problem. Then*

$$-(Ax^* - b) \in \partial(-g)(y)$$

if $f(x^*) + \langle y, Ax^* - b \rangle = g(y)$, i.e., if x^* minimizes $f(x) + \langle y, Ax - b \rangle$ over $x \in \mathbb{R}^d$.

Lemma 5.1 indicates that, given $y \in \mathbb{R}^p$, if we can find a minimizer x^* of the Lagrangian $f(x) + \langle y, Ax - b \rangle$, then a subgradient of $-g$ at y is given by $Ax^* - b$. As a result, we may regard the following iterations

$$\begin{aligned} x(t+1) &\in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} f(x) + \langle y(t), Ax - b \rangle \\ y(t+1) &= y(k) + \eta_t (Ax(t+1) - b) \end{aligned} \tag{5.3}$$

as performing the subgradient descent method on the function $(-g)$. Then, if strong duality holds for the convex optimization problem (5.2), as long as $f(x) + \langle y(t), Ax - b\rangle$ has a minimizer over $x \in \mathbb{R}^d$ for all t and $y(t)$ converges to an optimal dual variable, $f(x(t))$ will converge to the optimal value of (5.2). The algorithm (5.3) is called the *dual ascent* method.

It can be shown that, if f is μ-strongly convex, then a minimizer of $f(x) + \langle y, Ax - b\rangle$ over $x \in \mathbb{R}^d$ always exists, and $-g$ in this case is in fact $\|A\|_2^2/\mu$-smooth; if f is further L-smooth, then $-g$ will be $\|A\|_2^2/L$-strongly convex.[1] However, it seems tricky to find weaker conditions than strong convexity that guarantee the existence of a minimizer of $f(x) + \langle y(t), Ax - b\rangle$ for all t, and there are many scenarios where even those weaker conditions do not hold. Also, when f is not strongly convex, the dual objective function g may not be smooth, and we may need to use a diminishing sequence of step sizes η_t to guarantee the convergence of dual ascent as $t \to \infty$, which may slow down convergence. To overcome these shortcomings, the *method of multipliers* has been proposed:

$$\begin{aligned} x(t+1) &\in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \mathscr{L}_\rho(x, y(t)), \\ y(t+1) &= y(t) + \eta(Ax(t+1) - b), \end{aligned}$$

with the augmented Lagrangian defined as

$$L_\rho(x, y) = f(x) + \langle y, Ax - b\rangle + \frac{\rho}{2}\|Ax - b\|^2.$$

The following lemma shows that, under relatively mild conditions, we can guarantee that $\mathscr{L}_\rho(x, y)$ has a minimizer over $x \in \mathbb{R}^d$ for every y.

Lemma 5.2. *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function with a nonempty and closed epigraph, and let $A \in \mathbb{R}^{p\times d}$ and $b \in \mathbb{R}^p$. Suppose the relative interior of the set*

$$\left\{z \in \mathbb{R}^d : f^*(z) < +\infty\right\}$$

is nonempty and intersects with $\operatorname{range}(A^\mathsf{T})$*, where f^* denotes the convex conjugate of f. Then for any $y \in \mathbb{R}^d$ and $\rho > 0$, the set*

$$\operatorname*{arg\,min}_{x \in \mathbb{R}^d} \left(f(x) + \langle y, Ax - b\rangle + \frac{\rho}{2}\|Ax - b\|^2\right)$$

is nonempty.

Proof. Since the epigraph of f is closed, it is known from convex analysis that the convex conjugate of f^* is f [Rockafellar, 1970, Theorem 12.2]. Now consider the optimization problem

$$\begin{aligned} \min_{\mu \in \mathbb{R}^p, \nu \in \mathbb{R}^n} \quad & f^*(\rho\nu) + \langle \mu, y - \rho b\rangle + \frac{\rho}{2}\|\mu\|^2 \\ \text{s.t.} \quad & \rho A^\mathsf{T}\mu = \rho\nu. \end{aligned} \tag{5.4}$$

[1] This conclusion is a consequence of the fact that i) when a convex function f has a nonempty and closed epigraph, for any (x, z), we have $z \in \partial f(x)$ if and only if $x \in \partial f^*(z)$; and that ii) $\partial(-g)(y) = A\partial f^*(-A^\mathsf{T}y)$ when $f^*(z) < +\infty$ for all z. See [Rockafellar, 1970, Theorems 23.5 and 23.9].

Since the relative interior of $\mathrm{dom}(f^*) = \{z : f^*(z) < +\infty\}$ is nonempty and intersects with $\mathrm{range}(A^\mathsf{T})$, we can see that the above optimization problem satisfies Slater's condition. The objective function of the dual problem of (5.4) is given by

$$\begin{aligned}
&\inf_{\mu,\nu} \Big(f^*(\rho\nu) + \langle \mu, y - \rho b\rangle + \frac{\rho}{2}\|\mu\|^2 + \rho\langle x, A^\mathsf{T}\mu - \nu\rangle \Big) \\
=\ &\inf_{\nu} \left(f^*(\rho\nu) - \langle x, \rho\nu\rangle \right) - \frac{\rho}{2}\left\| Ax - b + y/\rho \right\|^2 \\
=\ &-f(x) - \frac{\rho}{2}\left\| Ax - b + y/\rho \right\|^2 \\
=\ &-\Big(f(x) + \langle y, Ax - b\rangle + \frac{\rho}{2}\|Ax - b\|^2 \Big) - \frac{\|y\|^2}{2\rho}.
\end{aligned}$$

Therefore the dual problem has a maximizer if and only if $f(x) + \langle y, Ax - b\rangle + \frac{\rho}{2}\|Ax - b\|^2$ has a minimizer. Finally, note that the optimization problem (5.4) can be reformulated as $\min_\mu f^*(A^\mathsf{T}\mu) + \langle \mu, y - \rho b\rangle + \rho\|\mu\|^2/2$, whose objective function is ρ-strongly convex. Thus (5.4) has a finite optimal value, and by Slater's condition, we see that an optimal dual variable exists, implying that $f(x) + \langle y, Ax - b\rangle + \frac{\rho}{2}\|Ax - b\|^2$ has a minimizer. □

The following theorem shows that, the method of multipliers converges with a properly chosen constant step size:

Theorem 5.1. *Suppose the conditions in Lemma 5.2 holds, and further suppose the problem* (5.2) *has a saddle point* (x^*, y^*)*. Then, by setting* $\eta = \rho$*, the iterates generated by the method of multipliers satisfy*

$$\lim_{t\to\infty} f(x(t)) = f(x^*) \qquad \text{and} \qquad \lim_{t\to\infty} Ax(t) - b = 0.$$

Note that in Theorem 5.1 we choose the step size to be $\eta = \rho$. One nice property of this choice is that, since $x(t+1)$ minimizes $\mathscr{L}_\rho(x, y(t))$, we get

$$\begin{aligned}
0 &\in \partial_x \mathscr{L}_\rho(x(t+1), y(t)) \\
&= \partial f(x(t+1)) + A^\mathsf{T}(y(t) + \rho(Ax(t+1) - b)) \\
&= \partial f(x(t+1)) + A^\mathsf{T} y(t+1) \\
&= \partial_x \mathscr{L}(x(t+1), y(t+1)),
\end{aligned}$$

where $\mathscr{L}$ is the ordinary Lagrangian. This indicates that the primal-dual pair $(x(t+1), y(t+1))$ will always be dual feasible. On the other hand, we mention that the method of multipliers will still converge when $\eta \in (0, 2\rho)$; see [Ryu and Yin, 2022, Chapter 8] for more details.

Now let's consider the optimization problem (5.1). Applying the method of multipliers to (5.1) gives

$$\begin{aligned}
(x_1(t+1), x_2(t+1)) = \ &\operatorname*{arg\,min}_{x_1,x_2} \Big(f_1(x_1) + f_2(x_2) + \langle y, A_1x_1 + A_2x_2 - b\rangle \\
&\qquad + \frac{\rho}{2}\|A_1x_1 + A_2x_2 - b\|^2 \Big), \\
y(t+1) = \ &y(t) + \eta(A_1x_1(t+1) + A_2x_2(t+1) - b).
\end{aligned}$$

By comparing it with ADMM, we can see that in the method of multipliers, the optimization over x_1 and x_2 is carried out jointly, while in ADMM, the optimization over x_1 and x_2 is carried out sequentially, and in the optimization of x_2 we plug in the updated value of x_1. This difference accounts for the term *alternating direction* in the name of ADMM.

Convergence of ADMM

The following theorem establishes the convergence of ADMM under relatively mild conditions:

Theorem 5.2. *Suppose the problem* (5.1) *has a saddle point* (x_1^*, x_2^*, y^*)*. Furthermore, suppose for each* $i = 1, 2$*, the relative interior of the set*

$$\left\{z \in \mathbb{R}^{d_i} : f_i^*(z) < +\infty\right\}$$

is nonempty and intersects with $\mathrm{range}(A_i^\mathsf{T})$*, where* f_i^* *denotes the convex conjugate of* f_i*. Then, the subproblems of ADMM for updating* $x_1(t)$ *and* $x_2(t)$ *always have solutions. Moreover, by choosing* $\eta = \rho$*, we have*

$$\lim_{t\to\infty} \left(f_1(x_1(t)) + f_2(x_2(t))\right) = f_1(x_1^*) + f_2(x_2^*), \quad \textit{and} \quad \lim_{t\to\infty} \left(A_1 x_1(t) + A_2 x_2(t) - b\right) = 0.$$

The conclusion in Theorem 5.2 that the subproblems for updating $x_1(t)$ and $x_2(t)$ always have solutions can be justified by Lemma 5.2, and we postpone the proof of convergence to Appendix 5.A. The convergence proof employs the following Lyapunov function

$$V(t) = \frac{\rho}{2}\|A_2(x_2(t) - x_2^*)\|^2 + \frac{1}{2\rho}\|y(t) - y^*\|^2,$$

from which we can intuitively interpret the parameter ρ as controlling the balance between primal and dual convergence: Large ρ facilitates convergence of the primal error $\|A_2(x_2(t) - x_2^*)\|$, while small ρ promotes convergence of the dual error $\|y(t) - y^*\|$.

Note that in Theorem 5.2 we let $\eta = \rho$, but it can be shown that we can relax this condition to be $\eta \in \left(0, \frac{\sqrt{5}+1}{2}\rho\right)$ and ADMM will still converge; see [Ryu and Yin, 2022, Chapter 8] for details. Also note that Theorem 5.1 can be viewed as a special case of Theorem 5.2.

Some Simple Examples

The main reason we employ sequential optimization instead of joint optimization in ADMM is that, in many scenarios, optimizing over x_1 and x_2 separately is usually easier than optimizing over (x_1, x_2) jointly, especially when the objective function is decomposed appropriately.

Example 5.1. Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two closed convex sets in $\mathbb{R}^d$ and suppose $\mathcal{X}_1 \cap \mathcal{X}_2 \neq \varnothing$. Consider the problem of finding $\mathcal{P}_{\mathcal{X}_1 \cap \mathcal{X}_2}[z]$ for any $z \in \mathbb{R}^d \backslash (\mathcal{X}_1 \cap \mathcal{X}_2)$, given that the projections $\mathcal{P}_{\mathcal{X}_1}$ and $\mathcal{P}_{\mathcal{X}_2}$ can be computed efficiently. This problem can be reformulated as an optimization

problem:

$$\begin{aligned} \min_{x_1,x_2\in\mathbb{R}^d} \quad & \frac{1}{2}\|z-x_1\|^2+\delta_{\mathcal{X}_1}(x_1)+\delta_{\mathcal{X}_2}(x_2) \\ \text{s.t.} \quad & x_1-x_2=0, \end{aligned}$$

where $\delta_{\mathcal{X}}$ denotes the indicator function of the convex set $\mathcal{X}$ defined by

$$\delta_{\mathcal{X}}(x)=\begin{cases}0, & x\in\mathcal{X},\\ +\infty, & x\notin\mathcal{X}.\end{cases}$$

The augmented Lagrangian is

$$\mathscr{L}_\rho(x_1,x_2,y)=\frac{1}{2}\|z-x_1\|^2+\delta_{\mathcal{X}_1}(x_1)+\delta_{\mathcal{X}_2}(x_2)+\langle y,x_1-x_2\rangle+\frac{\rho}{2}\|x_1-x_2\|^2.$$

Optimizing $\mathscr{L}_\rho(x_1,x_2,y)$ over (x_1,x_2) jointly seems not easy to handle. On the other hand, by applying ADMM, we get the following iterations

$$\begin{aligned} x_1(t+1) &= \arg\min_{x_1}\mathscr{L}_\rho(x_1,x_2(t),y(t)) \\ &= \arg\min_{x_1}\left(\frac{1+\rho}{2}\left\|x_1-\frac{z+\rho x_2(t)-y(t)}{1+\rho}\right\|^2+\delta_{\mathcal{X}_1}(x_1)\right) \\ &= \mathcal{P}_{\mathcal{X}_1}\left[\frac{z+\rho x_2(t)-y(t)}{1+\rho}\right], \\ x_2(t+1) &= \arg\min_{x_2}\mathscr{L}_\rho(x_1(t+1),x_2,y(t)) \\ &= \arg\min_{x_2}\left(\frac{\rho}{2}\left\|x_2-\frac{y(t)+\rho x_1(t+1)}{\rho}\right\|^2+\delta_{\mathcal{X}_2}(x_2)\right) \\ &= \mathcal{P}_{\mathcal{X}_2}\left[x_1(t+1)+\frac{y(t)}{\rho}\right], \\ y(t+1) &= y(t)+\rho(x_1(t+1)-x_2(t+1)). \end{aligned}$$

Theorem 5.2 together with the result in Exercise 5.1 guarantees that $x_1(t)$ and $x_2(t)$ generated by the above iterations converge to $\mathcal{P}_{\mathcal{X}_1\cap\mathcal{X}_2}[z]$. □

Exercise 5.1. Let $\mathcal{X}_1,\mathcal{X}_2\subseteq\mathbb{R}^d$ be two closed and convex sets with a nonempty intersection. Let $z\in\mathbb{R}^d\backslash(\mathcal{X}_1\cap\mathcal{X}_2)$ be arbitrary, and suppose $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ are two sequences satisfying

1. $u_n\in\mathcal{X}_1$ and $v_n\in\mathcal{X}_2$ for all $n\in\mathbb{N}$;
2. $\|u_n-v_n\|\to 0$ as $n\to\infty$;
3. $\|u_n-z\|\to\|\mathcal{P}_{\mathcal{X}_1\cap\mathcal{X}_2}[z]-z\|$ as $n\to\infty$.

Show that $u_n\to\mathcal{P}_{\mathcal{X}_1\cap\mathcal{X}_2}[z]$ as $n\to\infty$. □

Example 5.2. Consider the LASSO problem formulated as

$$\min_{x} \quad \frac{1}{2}\|Ax-b\|^2+\lambda\|x\|_1,$$

where $\lambda>0$ is a regularization parameter. We reformulate this problem as

$$\begin{aligned}\min_{x_1,x_2} \quad & \frac{1}{2}\|Ax_1-b\|^2+\lambda\|x_2\|_1\\ \text{s.t.} \quad & x_1=x_2.\end{aligned}$$

The corresponding augmented Lagrangian is then

$$\mathscr{L}_\rho(x_1,x_2,y)=\frac{1}{2}\|Ax_1-b\|^2+\lambda\|x_2\|_1+\langle y,x_1-x_2\rangle+\frac{\rho}{2}\|x_1-x_2\|^2.$$

The ADMM iterations are given by

$$\begin{aligned}x_1(t+1) &= \arg\min_{x_1}\mathscr{L}_\rho(x_1,x_2(t),y(t))\\ &= \arg\min_{x_1}\left(\frac{1}{2}x_1^{\mathsf{T}}(A^{\mathsf{T}}A+\rho I)x_1+\langle y(t)-A^{\mathsf{T}}b-\rho x_2(t),x_1\rangle\right)\\ &= \left(A^{\mathsf{T}}A+\rho I\right)^{-1}\left(A^{\mathsf{T}}b+\rho x_2(t)-y(t)\right)\\ x_2(t+1) &= \arg\min_{x_2}\mathscr{L}_\rho(x_1(t+1),x_2,y(t))\\ &= \arg\min_{x_2}\sum\nolimits_i\left(\lambda|x_{2,i}|+\frac{\rho}{2}\left|x_{2,i}-\frac{y_i+\rho x_{1,i}(t+1)}{\rho}\right|^2\right)\\ &= S_{\lambda/\rho}\left[x_1(t+1)+\frac{y(t)}{\rho}\right],\\ y(t+1) &= y(t)+\rho(x_1(t+1)-x_2(t+1)),\end{aligned}$$

where the *soft thresholding* operator $\mathcal{S}_\kappa$ is defined by

$$(\mathcal{S}_\kappa[v])_i\coloneqq\arg\min_{x_i}\left(\kappa|x_i|+\frac{1}{2}|x_i-v_i|^2\right)=\begin{cases}v_i-\kappa, & v_i>\kappa,\\ 0, & v_i\in[-\kappa,\kappa],\\ v_i+\kappa, & v_i<-\kappa.\end{cases}$$

Note that the update of x_1 is essentially a *ridge regression* step, which is usually the main contributor to the total complexity of the algorithm, while the update of x_2 is a simple soft thresholding step, whose computation complexity is in general negligible compared to the update of x_1. □

5.2 Decentralized ADMM for Consensus Optimization

In this section, we show how to design a distributed optimization algorithm for consensus optimization based on ADMM.

Consider a group of N agents connected by a communication network. Recall that the consensus optimization problem is formulated as

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

where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is a local cost function associated with agent i. We assume that each f_i is a convex function, and also assume that the topology of the communication network can be described by an undirected connected graph $\mathcal{G} = (\{1,\dots,N\}, \mathcal{E})$.

It's not hard to see that, since the graph $\mathcal{G}$ is connected, the consensus optimization problem can be reformulated as

$$\begin{aligned}\min_{\substack{x_1,\dots,x_N\in\mathbb{R}^d,\\ z_e\in\mathbb{R}^d: e\in\mathcal{E}}} \ & \sum_{i=1}^N f_i(x_i)\\ \text{s.t.} \ \ & x_i = z_e,\ x_j = z_e, \qquad \forall e=\{i,j\}\in\mathcal{E}.\end{aligned}$$

By setting $x = (x_1,\dots,x_N)$, $z = (z_e : e\in\mathcal{E})$, $f(x) = \sum_{i=1}^N f_i(x_i)$, $g(z) = 0$, we can see that the above formulation can be written abstractly as

$$\begin{aligned}\min_{x,z} \ & f(x) + g(z)\\ \text{s.t.} \ & Ax + Bz = 0\end{aligned}$$

for some matrices A and B, showing that the formulation fits the problem setup of ADMM. The corresponding augmented Lagrangian is given by

$$\begin{aligned}\mathscr{L}_\rho(x,z,y) = & \sum_{i=1}^N f_i(x_i) + \sum_{e=\{i,j\}\in\mathcal{E}} \left(\langle y_{e,i}, x_i - z_e\rangle + \langle y_{e,j}, x_j - z_e\rangle\right)\\ & + \frac{\rho}{2}\sum_{e=\{i,j\}\in\mathcal{E}} \left(\|x_i - z_e\|^2 + \|x_j - z_e\|^2\right),\end{aligned}$$

where $y = ((y_{e,i}, y_{e,j}) : e = \{i,j\}\in\mathcal{E})$. The x-update of ADMM is given by

$$\begin{aligned}x(t+1) = & \ \arg\min_x \Bigg\{ \sum_{i=1}^N f_i(x_i)\\ & + \sum_{e=\{i,j\}} \left[\langle y_{e,i}(t) - \rho z_e(t), x_i\rangle + \langle y_{e,j}(t) - \rho z_e(t), x_j\rangle + \frac{\rho}{2}(\|x_i\|^2 + \|x_j\|^2)\right]\Bigg\}\\ = & \ \arg\min_x \sum_{i=1}^N \left(f_i(x_i) + \sum_{e\in\mathcal{E}: e\ni i} \langle y_{e,i}(t) - \rho z_e(t), x_i\rangle + \frac{\rho\deg(i)}{2}\|x_i\|^2\right).\end{aligned}$$

It's not hard to see that the objective function on right-hand side above is separable. Therefore the x-update can be written as

$$x_i(t+1) = \ \arg\min_{x_i\in\mathbb{R}^d} \left(f_i(x_i) + \sum_{e\in\mathcal{E}: e\ni i} \langle y_{e,i}(t) - \rho z_e(t), x_i\rangle + \frac{\rho\deg(i)}{2}\|x_i\|^2\right).$$

The z-update of ADMM is given by

$$z(t+1) = \arg\min_{z} \sum_{e=\{i,j\}} \Big[-\langle y_{e,i}(t) + y_{e,j}(t) + \rho x_i(t+1) + \rho x_j(t+1), z_e\rangle + \frac{\rho}{2}\left(\|z_e\|^2 + \|z_e\|^2\right)\Big].$$

The right-hand side can be solve explicitly, which leads to

$$z_e(t+1) = \frac{\rho^{-1}(y_{e,i}(t) + y_{e,j}(t)) + x_i(t+1) + \rho x_j(t+1)}{2}, \qquad e = \{i,j\}.$$

Finally, the y-update of ADMM is given by

$$y_{e,i}(t+1) = y_{e,i}(t) + \rho(x_i(t+1) - z_e(t+1)).$$

We summarize the ADMM iterations as follows:

$$\begin{aligned}
x_i(t+1) &= \arg\min_{x_i\in\mathbb{R}^d} \left(f_i(x_i) + \sum_{e\in\mathcal{E}: e\ni i} \langle y_{e,i}(t) - \rho z_e(t), x_i\rangle + \frac{\rho\deg(i)}{2}\|x_i\|^2 \right), \\
z_e(t+1) &= \frac{\rho^{-1}(y_{e,i}(t) + y_{e,j}(t)) + x_i(t+1) + x_j(t+1)}{2}, \\
y_{e,i}(t+1) &= y_{e,i}(t) + \rho(x_i(t+1) - z_e(t+1)), \qquad\qquad e = \{i,j\}.
\end{aligned}$$

We then try to simplify the above iterations by eliminating some variables. First, note that by plugging the z-update rule into the y-update rule, we get

$$\begin{aligned}
y_{e,i}(t+1) &= y_{e,i}(t) + \rho x_i(t+1) - \frac{1}{2}\left(y_{e,i}(t) + y_{e,j}(t) + \rho x_i(t+1) + \rho x_j(t+1)\right) \\
&= \frac{1}{2}\left(y_{e,i}(t) - y_{e,j}(t) + \rho(x_i(t+1) - x_j(t+1))\right), \qquad e = \{i,j\}.
\end{aligned}$$

By exchanging i and j and taking the summation, we get $y_{e,i}(t+1) + y_{e,j}(t+1) = 0$ for all $t \geq 0$. Without loss of generality we may adopt the initialization $y_{e,i}(0) = y_{e,j} = 0$ for all $e = \{i,j\} \in \mathcal{E}$, and then we get

$$z_e(t+1) = \frac{1}{2}(x_i(t+1) + x_j(t+1)).$$

We then elimiate z_e in the x-update rule and obtain

$$x_i(t+1) = \arg\min_{x_i\in\mathbb{R}^d} \left(f_i(x_i) + \sum_{e\in\mathcal{E}: e=\{i,j\}} \left\langle y_{e,i}(t) - \frac{\rho}{2}\left(x_i(t) + x_j(t)\right), x_i\right\rangle + \frac{\rho\deg(i)}{2}\|x_i\|^2 \right).$$

To further simplify the iterations, we introduce the variable

$$v_i(t) = \frac{1}{\deg(i)} \sum_{e\in\mathcal{E}: e=\{i,j\}} \left[-\rho^{-1} y_{e,i}(t) + \frac{1}{2}\left(x_i(t) + x_j(t)\right)\right].$$

Then the x-update can be written as

$$\begin{aligned}
x_i(t+1) &= \arg\min_{x_i\in\mathbb{R}^d} \left(f_i(x_i) - \rho\deg(i)\langle v_i(t), x_i\rangle + \frac{\rho\deg(i)}{2}\|x_i\|^2 \right) \\
&= \arg\min_{x_i\in\mathbb{R}^d} \left(f_i(x_i) + \frac{\rho\deg(i)}{2}\|x_i - v_i(t)\|^2 \right).
\end{aligned}$$

Moreover,

$$
\begin{aligned}
& v_i(t+1) - v_i(t) \\
&= \frac{1}{\deg(i)} \sum_{e \in \mathcal{E}: e=\{i,j\}} \left[-\rho^{-1}(y_{e,i}(t+1) - y_{e,i}(t)) + \frac{1}{2}(x_i(t+1) + x_j(t+1) - x_i(t) - x_j(t)) \right] \\
&= \frac{1}{\deg(i)} \sum_{e \in \mathcal{E}: e=\{i,j\}} \left[-\frac{1}{2}(y_{e,i}(t) + y_{e,j}(t)) + x_j(t+1) - \frac{1}{2}(x_i(t) + x_j(t)) \right] \\
&= \frac{1}{\deg(i)} \sum_{j \in \mathcal{N}_i} \left(x_j(t+1) - \frac{1}{2} x_j(t) \right) - \frac{1}{2} x_i(t),
\end{aligned}
$$

where in the second step we plugged in the y-update rule, and in the third step we used $y_{e,i}(t) + y_{e,j}(t) = 0$. Summarizing the previous derivations, we get

$$
\begin{aligned}
x_i(t+1) &= \operatorname*{arg\,min}_{x_i \in \mathbb{R}^d} \left(f_i(x_i) + \frac{\rho \deg(i)}{2} \|x_i - v_i(t)\|^2 \right), \\
v_i(t+1) &= v_i(t) + \frac{1}{\deg(i)} \sum_{j \in \mathcal{N}_i} x_j(t+1) - \frac{1}{2} \sum_{j \in \mathcal{N}_i} x_j(t) - \frac{1}{2} x_i(t).
\end{aligned}
$$

It's not hard to see that the above iterations can be implemented in a decentralized fashion by the group of agents via the communication network.

Notes on References

One of the standard references on ADMM is the review paper [Boyd et al., 2011], and we highly recommend interested readers to go through the various examples and applications of ADMM presented in this paper. However, it should be noted that the paper mistakenly claimed that the subproblems for finding $x_1(t+1)$ and $x_2(t+1)$ have solutions as long as f_1 and f_2 have nonempty closed epigraphs; see also [Chen et al., 2017].

Many technical materials in this chapter are adapted from [Ryu and Yin, 2022], including Lemma 5.2, Theorem 5.1, Theorem 5.2, and the decentralized ADMM algorithm. The book [Ryu and Yin, 2022] also presents some variations and extensions of ADMM.

The analysis in this chapter only establishes the convergence of ADMM but without bounding the convergence rate. Some existing works that analyze the convergence rate of ADMM include [He and Yuan, 2012], [Shi et al., 2014], [Deng and Yin, 2016], [Davis and Yin, 2017], etc. We also refer to the Bibliographical Notes of [Ryu and Yin, 2022, Chapter 8], which provides an excellent survey of the history and relevant works of ADMM.

5.A Proof of Convergence of ADMM

First of all, note that

$$\begin{aligned}&\mathscr{L}(x_1(t+1), x_2(t+1), y^*) - \mathscr{L}(x_1^*, x_2^*, y^*)\\ &= f_1(x_1(t+1)) - f_1(x_1^*) + f_2(x_2(t+1)) - f_2(x_2^*) + \langle y^*, A_1x_1(t+1) + A_2(x_2(t+1) - b)\rangle,\end{aligned} \tag{5.5}$$

where we employ the equality $A_1x_1^* + A_2x_2^* = b$. To bound the differences $f_1(x_1(t+1)) - f_1(x_1^*)$ and $f_2(x_2(t+1)) - f_2(x_2^*)$, we utilize the update rules of $x_1(t)$ and $x_2(t)$. Specifically, $x_1(t+1) \in \arg\min_{x_1} \mathscr{L}_\rho(x_1, x_2(t), y(t))$ implies

$$\begin{aligned}0 &\in \partial_{x_1}\mathscr{L}_\rho(x_1(t+1), x_2(t), y(t))\\ &= \partial f_1(x_1(t+1)) + A_1^\mathsf{T} y(t) + \rho A_1^\mathsf{T}(A_1x_1(t+1) + A_2x_2(t) - b),\end{aligned}$$

i.e., $-A_1^\mathsf{T}(y(t) + \rho(A_1x_1(t+1) + A_2x_2(t) - b)) \in \partial f_1(x_1(t))$. Therefore, for an arbitrary $x_1 \in \mathbb{R}^{d_1}$, we have

$$\begin{aligned}f_1(x_1) &\geq f_1(x_1(t+1)) - \langle A_1^\mathsf{T}(y(t) + \rho(A_1x_1(t+1) + A_2x_2(t) - b)), x_1 - x_1(t+1)\rangle\\ &= f_1(x_1(t+1)) + \langle y(t+1) - \rho A_2(x_2(t+1) - x_2(t)), A_1(x_1(t+1) - x_1)\rangle\end{aligned} \tag{5.6}$$

Similarly, from the update rule of $x_2(t)$, we get

$$\begin{aligned}f_2(x_2) &\geq f_2(x_2(t+1)) - \langle A_2^\mathsf{T}(y(t) + \rho(A_1x_1(t+1) + A_2x_2(t+1) - b)), x_2 - x_2(t+1)\rangle\\ &= f_2(x_2(t+1)) + \langle y(t+1), A_2(x_2(t+1) - x_2)\rangle.\end{aligned} \tag{5.7}$$

By setting $x_1 = x_1^*$ in (5.6) and $x_2 = x_2^*$ in (5.7) and plugging the two inequalities into (5.5), we get

$$\begin{aligned}&\mathscr{L}(x_1(t+1), x_2(t+1), y^*) - \mathscr{L}(x_1^*, x_2^*, y^*)\\ &\leq \langle y^* - y(t+1), A_1x_1(t+1) + A_2x_2(t+1) - b)\rangle\\ &\quad + \rho\langle A_2(x_2(t+1) - x_2(t)), A_1(x_1(t+1) - x_1^*)\rangle.\end{aligned}$$

We then plug in

$$A_1(x_1(t+1) - x_1^*) = -A_2(x_2(t+1) - x_2^*) + \rho^{-1}(y(t+1) - y(t))$$

and

$$\langle y^* - y(t+1), A_1x_1(t+1) + A_2x_2(t+1) - b\rangle = \rho^{-1}\langle y^* - y(t+1), y(t+1) - y(t)\rangle$$

to get

$$\begin{aligned}&\mathscr{L}(x_1(t+1), x_2(t+1), y^*) - \mathscr{L}(x_1^*, x_2^*, y^*)\\ &\leq -\frac{1}{\rho}\langle y(t+1) - y^*, y(t+1) - y(t)\rangle - \rho\langle A_2(x_2(t+1) - x_2(t)), A_2(x_2(t+1) - x_2^*)\rangle\\ &\quad + \langle A_2(x_2(t+1) - x_2(t)), y(t+1) - y(t)\rangle\end{aligned} \tag{5.8}$$

To further bound the inner product $\langle A_2(x_2(t+1) - x_2(t)), y(t+1) - y(t)\rangle$, we set $x_2 = x_2(t)$ in (5.7) and get

$$f_2(x_2(t)) \geq f_2(x_2(t+1)) + \langle y(t+1), A_2(x_2(t+1) - x_2(t))\rangle.$$

Then, we decrement the index $t+1$ to t in (5.7) and set $x_2 = x_2(t+1)$ to get

$$f_2(x_2(t+1)) \geq f_2(x_2(t)) + \langle y(t), A_2(x_2(t) - x_2(t+1))\rangle.$$

Summing the above two inequalities leads to

$$0 \geq \langle y(t+1) - y(t), A_2(x_2(t+1) - x_2(t))\rangle,$$

and together with (5.8) we get

$$\begin{aligned}
&\mathscr{L}(x_1(t+1), x_2(t+1), y^*) - \mathscr{L}(x_1^*, x_2^*, y^*) \\
\leq\ & -\frac{1}{\rho}\langle y(t+1) - y^*, y(t+1) - y(t)\rangle - \rho\langle A_2(x_2(t+1) - x_2(t)), A_2(x_2(t+1)) - x_2^*\rangle.
\end{aligned}$$

We now introduce the Lyapunov function

$$V(t) = \frac{\rho}{2}\|A_2(x_2(t) - x_2^*)\|^2 + \frac{1}{2\rho}\|y(t) - y^*\|^2.$$

Then it can be seen that

$$\begin{aligned}
V(t) &= \frac{\rho}{2}\|A_2(x_2(t) - x_2(t+1) + x_2(t+1) - x_2^*)\|^2 + \frac{1}{2\rho}\|y(t) - y(t+1) + y(t+1) - y^*\|^2 \\
&= V(t+1) - \rho\langle A_2(x_2(t+1) - x_2(t)), A_2(x_2(t+1) - x_2^*)\rangle + \frac{\rho}{2}\|A_2(x_2(t+1) - x_2(t))\|^2 \\
&\quad - \frac{1}{\rho}\langle y(t+1) - y(t), y(t+1) - y^*\rangle + \frac{1}{2\rho}\|y(t+1) - y(t)\|^2 \\
&\geq V(t+1) + \frac{\rho}{2}\|A_2(x_2(t+1) - x_2(t))\|^2 + \frac{1}{2\rho}\|y(t+1) - y(t)\|^2 \\
&\quad + \mathscr{L}(x_1(t+1), x_2(t+1), y^*) - \mathscr{L}(x_1^*, x_2^*, y^*),
\end{aligned}$$

or

$$\begin{aligned}
V(t+1) \leq V(t) - &\left(\frac{\rho}{2}\|A_2(x_2(t+1) - x_2(t))\|^2 + \frac{1}{2\rho}\|y(t+1) - y(t)\|^2\right) \\
&- (\mathscr{L}(x_1(t+1), x_2(t+1), y^*) - \mathscr{L}(x_1^*, x_2^*, y^*)).
\end{aligned}$$

Note that $\mathscr{L}(x_1(t+1), x_2(t+1), y^*) - \mathscr{L}(x_1^*, x_2^*, y^*) \geq 0$ since (x_1^*, x_2^*) minimizes $\mathscr{L}(x_1, x_2, y^*)$. By taking the telescoping sum and noting that $V(t) \geq 0$ for all t, we get

$$\begin{aligned}
V(0) \geq &\sum_{t=0}^{\infty}\left(\frac{\rho}{2}\|A_2(x_2(t+1) - x_2(t))\|^2 + \frac{1}{2\rho}\|y(t+1) - y(t)\|^2\right) \\
&+ \sum_{t=1}^{\infty}(\mathscr{L}(x_1(t), x_2(t), y^*) - \mathscr{L}(x_1^*, x_2^*, y^*)),
\end{aligned}$$

which implies that $A_1x_1(t) + A_2x_2(t) - b = \rho^{-1}(y(t) - y(t-1)) \to 0$ and

$$\begin{aligned}
&f_1(x_1(t)) + f_2(x_2(t)) - f_1(x_1^*) - f_2(x_2^*) \\
&= \mathscr{L}(x_1(t), x_2(t), y^*) - \mathscr{L}(x_1^*, x_2^*, y^*) - \langle y^*, A_1x_1(t) + A_2x_2(t) - b\rangle \\
&\to 0
\end{aligned}$$

as $t \to \infty$.

Bibliography

[Boyd et al., 2011] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

[Chen et al., 2017] Chen, L., Sun, D., and Toh, K.-C. (2017). A note on the convergence of ADMM for linearly constrained convex optimization problems. *Computational Optimization and Applications*, 66:327–343.

[Davis and Yin, 2017] Davis, D. and Yin, W. (2017). Faster convergence rates of relaxed Peaceman–Rachford and ADMM under regularity assumptions. *Mathematics of Operations Research*, 42(3):783–805.

[Deng and Yin, 2016] Deng, W. and Yin, W. (2016). On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66:889–916.

[He and Yuan, 2012] He, B. and Yuan, X. (2012). On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709.

[Rockafellar, 1970] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

[Ryu and Yin, 2022] Ryu, E. K. and Yin, W. (2022). *Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators*. Cambridge University Press, Cambridge, UK.

[Shi et al., 2014] Shi, W., Ling, Q., Yuan, K., Wu, G., and Yin, W. (2014). On the linear convergence of the ADMM in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761.

Chapter 6

Distributed Averaging and Optimization over Time-Varying Networks

6.1 Time-Varying Communication Networks

In many practical scenarios, the communication network connecting the group of agents is not static and varies with time. For example, consider a group of N drones that are equipped with wireless communication modules. Due to limited power supply, the wireless communication module only allows the drone to communicate with other drones within a limited physical range. Specifically, at time t, suppose the wireless signal sent by drone i can be received by drone j only when $\|\mathbf{r}_j(t)-\mathbf{r}_i(t)\| \leq R_i$, where $\mathbf{r}_i$ denotes the position vector of drone i, and $R_i > 0$ represents the radius of the range that the wireless signal sent by drone i can cover. Then the topology of the wireless communication network at time t can be modeled by the directed graph $\mathcal{G}(t)$ defined by

$$\begin{aligned}&\mathcal{G}(t) = \{(1,\dots,N), \mathcal{E}(t)\},\\ &(i,j) \in \mathcal{E}(t) \iff \|\mathbf{r}_j(t) - \mathbf{r}_i(t)\| \leq R_i.\end{aligned}$$

It can be seen that, as the positions of the drones change, the communication graph $\mathcal{G}(t)$ in general will also change. Also, since each R_i can be different, the communication graph in general will be directed.

The time-varying nature of the communication graph imposes further challenges on the design and analysis of multi-agent distributed averaging and consensus optimization algorithms. From the previous chapters, we can see that one of the core components of many consensus methods is the weighted-sum iteration

$$x_i(t+1) = \sum_{i=1}^{N} W_{ij}(t) x_j(t),$$

where the weights $W_{ij}(t)$ are compatible with the topology of the communication network. Note that since the communication network is now time-varying, the weights will necessarily become dependent on t. As usual, we let $W(t) =\in \mathbb{R}^{N\times N}$ denote the weight matrix whose entries are the weights $W_{ij}(t)$. We also denote

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

The weighted-sum iteration can then be equivalently written as

$$\mathsf{X}(t+1) = W(t)\mathsf{X}(t),$$

which leads to

$$\mathsf{X}(t) = W(t-1)\cdots W(s+1)\cdot W(s)\mathsf{X}(s)$$

for any $s, t \in \mathbb{N}$ with $s > t$. Therefore, it can be expected that the core of our analysis should lie in characterizing the behavior of the product

$$W[s,t) \coloneqq W(t-1)\cdots W(s+1)\cdot W(s),$$

especially when $t-s$ is sufficiently large.

Before proceeding, we first list some useful notions and terminologies.

Definition 6.1. Let $A \in \mathbb{R}^{N\times N}$ be a real matrix.

1. A is said to be *row-stochastic*, if every entry of A is nonnegative, and $A\mathbf{1} = \mathbf{1}$.

   A is said to be *column-stochastic*, if A^T is row-stochastic.

   A is said to be *doubly stochastic*, if it is both row-stochastic and column-stochastic.
2. We say that $\mathcal{G} = (\{1,\dots,N\}, \mathcal{E})$ is the associated directed graph of A, if for all $i, j \in \{1,\dots,N\}$, we have $A_{ij} \neq 0 \iff (j,i) \in \mathcal{E}$.

Exercise 6.1. Let $A, A(1), \dots, A(N-1)$ be $N\times N$ real matrices with nonnegative entries, and suppose their associated directed graphs are all strongly connected. Moreover, suppose their diagonal elements are all positive.

1. Let $u \in \mathbb{R}^N$ be an arbitrary vector with at least one zero entry. Show that the number of nonzero entries of Au is strictly greater than the number of nonzero entries of u.
2. Show that all entries of $A(1)A(2)\cdots A(N-1)$ are positive.
3. Suppose there exists $\epsilon > 0$ such that all positive entries of $A(t)$ are lower bounded by ϵ for any $t = 1,\dots,N-1$. Show that any entry of $A(1)A(2)\cdots A(N-1)$ is greater than or equal to ϵ^{N-1}. □

We first consider the situation where the graph $\mathcal{G}(t)$ is strongly connected for all t. The following theorem provides a fundamental tool for analyzing the convergence behavior of $W[s,t)$.

Theorem 6.1. *For each $t \in \mathbb{N}$, let $W(t) \in \mathbb{R}^{N\times N}$ be a weight matrix with associated directed graph $\mathcal{G}(t) = (\{1,\dots,N\}, \mathcal{E}(t))$. Suppose the following conditions are satisfied:*

1. *Each $\mathcal{G}(t)$ is strongly connected.*
2. *Each $W(t)$ is row-stochastic.*
3. *$W_{ii}(t) > 0$ for all $i = 1,\dots,N$ and all $t \in \mathbb{N}$.*
4. *There exists $\epsilon > 0$ such that for any $t \in \mathbb{N}$, $W_{ij}(t) \geq \epsilon$ whenever $W_{ij}(t) > 0$.*

Then there exist $C > 0$, $\sigma \in (0,1)$, $\eta > 0$ and a sequence of vectors

$$w(s) \in \left\{ w \in \mathbb{R}^N : w_i \geq \eta,\ \forall i \text{ and } \mathbf{1}^\mathsf{T} w = 1 \right\},\ s = 0,1,2,\dots$$

such that

$$\left| \left(W[s,t]\right)_{ij} - w_j(s) \right| = \left| \left(W[s,t] - \mathbf{1}w(s)^\mathsf{T}\right)_{ij} \right| \leq C\sigma^{t-s}$$

for any i, j and any $t > s \geq 0$. The constants C, σ and η only depend on the number of agents N and the lower bound of the positive entries ϵ. Furthermore, $w(s) = \frac{1}{N}\mathbf{1}$ if each $W(t)$ is doubly-stochastic.

The proof of Theorem 6.1 is postponed to Appendix 6.A.

Theorem 6.1 lies at the core of the analysis of many distributed averaging/consensus optimization algorithms for time-varying communication networks. In the next section, we provide an example to show how Theorem 6.1 can be applied to analyze the extension of the radio-consensus method with time-varying communication networks.

Exercise 6.2. Under the setting of Theorem 6.1, show that the vectors $w(s)$, $s = 0,1,2,\dots$ satisfy

$$w(s)^\mathsf{T} = w(t)^\mathsf{T} W[s,t]$$

for all $t > s \geq 0$.

(Hint: Show the identity

$$\mathbf{1}\left(w(s)^\mathsf{T} - w(t)^\mathsf{T} W[s,t]\right) = \mathbf{1}w(s)^\mathsf{T} - W[s,\tau] + \left((W[t,\tau] - \mathbf{1}w(t)^\mathsf{T}\right) W[s,t]$$

for any $\tau > t$, and then use Theorem 6.1 to find the limit of the right-hand side as $\tau \to \infty$.) □

Exercise 6.3. Under the setting of Theorem 6.1, suppose there exists a sequence of vectors $\tilde{w}(s) \in \mathbb{R}^N$, $s = 0,1,2,\dots$ such that all entries of $\tilde{w}(s)$ are nonnegative, $\mathbf{1}^\mathsf{T}\tilde{w}(s) = 1$ and

$$\tilde{w}(s)^\mathsf{T} = \tilde{w}(t)^\mathsf{T} W[s,t]$$

for all $t > s \geq 0$. Let $w(s)$, $s = 0,1,2,\dots$ denote the vectors obtained from Theorem 6.1.

1. Show that
$$\tilde{w}(s)^{\mathsf{T}} - w(s)^{\mathsf{T}} = \tilde{w}(t)^{\mathsf{T}}(W[s,t) - \mathbf{1}w(s)^{\mathsf{T}})$$
for all $t > s \geq 0$.
2. Show that $\tilde{w}(s) = w(s)$ for all $s \in \mathbb{N}$. □

Exercise 6.4. Suppose we work under the setting of Theorem 6.1. Let $b \in \mathbb{R}^N$ and $t > s \geq 0$ be arbitrary, and let
$$a = W[s,t)b.$$
Show that
$$\left\|a - \mathbf{1}w(t)^{\mathsf{T}}a\right\| \leq CN\sigma^{t-s}\left\|b - \mathbf{1}w(s)^{\mathsf{T}}b\right\|,$$
where the constants C and σ are from Theorem 6.1. □

6.2 The Push-Sum Method for Distributed Averaging

Let $\mathcal{G}(t) = (\{1,\dots,\mathcal{E}(t)\})$ be a sequence of strongly connected directed graphs of which every node has a self-loop. Suppose each agent i is associated with a given vector x_i, and the goal of the group of agents is to find the average value $\frac{1}{N}\sum_{i=1}^N x_i$ via local computation and coordination through the time-varying communication network.

To design a distributed averaging algorithm for this problem setup, we consider adapting the ratio consensus method (2.9) to the time-varying setting:

$$\begin{aligned} y_i(t+1) &= \sum_{j=1}^N W_{ij}(t)y_j(t), \qquad y_i(0) = x_i, \\ z_i(t+1) &= \sum_{j=1}^N W_{ij}(t)z_j(t), \qquad z_i(0) = 1, \\ x_i(t) &= \frac{y_i(t)}{z_i(t)}. \end{aligned} \tag{6.1}$$

Here the weights are taken as
$$W_{ij}(t) = \begin{cases} \dfrac{1}{|\{k : (j,k) \in \mathcal{E}(t)\}|}, & (j,i) \in \mathcal{E}(t), \\ 0, & (j,i) \notin \mathcal{E}(t), \end{cases}$$
where $|\{k : (j,k) \in \mathcal{E}(t)\}|$ denotes the number of elements in the set $\{k : (j,k) \in \mathcal{E}(t)\}$, which is equal to the number of out-neighbors of node j (excluding j itself) plus 1. Note that we have $\frac{1}{N} \leq W_{ii}(t) \leq 1$ and $W(t)^{\mathsf{T}}\mathbf{1} = \mathbf{1}$ for every i, indicating that $W(t)$ is a *column-stochastic* matrix with positive diagonals. Moreover, $W_{ij}(t) \geq 1/N$ whenever $W_{ij}(t) > 0$.

The iterations (6.1) can be written in the following compact form:

$$\begin{aligned} \mathsf{Y}(t+1) &= W(t)\mathsf{Y}(t), \qquad \mathsf{Y}(0) = \begin{bmatrix} — \, x_1^\mathsf{T} \, — \\ \vdots \\ — \, x_N^\mathsf{T} \, — \end{bmatrix}, \\ z(t) &= W(t)z(t), \qquad z(0) = \mathbf{1}, \\ \mathsf{X}(t) &= Z(t)^{-1}\mathsf{Y}(t), \end{aligned}$$

where

$$\mathsf{Y}(t) \coloneqq \begin{bmatrix} — \, y_1(t)^\mathsf{T} \, — \\ \vdots \\ — \, y_N(t)^\mathsf{T} \, — \end{bmatrix}, \qquad \mathsf{X}(t) \coloneqq \begin{bmatrix} — \, x_1(t)^\mathsf{T} \, — \\ \vdots \\ — \, x_N(t)^\mathsf{T} \, — \end{bmatrix},$$

and

$$z(t) \coloneqq \begin{bmatrix} z_1(t) \\ \vdots \\ z_N(t) \end{bmatrix}, \qquad Z(t) \coloneqq \operatorname{diag}(z(t)) = \begin{bmatrix} z_1(t) & & \\ & \ddots & \\ & & z_N(t) \end{bmatrix}.$$

The iterations (6.1) are sometimes called the *push-sum* method.

In order for each $x_i(t)$ to be well defined for all t, we need to make sure that no entries of $z(t)$ become zero. This can be guaranteed by the following lemma:

Lemma 6.1. *For every $t \in \mathbb{N}$, we have*

$$\frac{1}{N^{N-2}} \le z_i(t) \le N$$

for all $i = 1, \dots, N$.

Exercise 6.5. In this exercise, you are asked to prove Lemma 6.1 by the following steps:

1. Show that every element of $z_i(t)$ is nonnegative and $\mathbf{1}^\mathsf{T} z_i(t) = N$ for each $t \in \mathbb{N}$. Then conclude that $z_i(t) \le N$ for all $i = 1, \dots, N$ and $t \in \mathbb{N}$.
2. By using $W_{ii}(t) \ge 1/N$, show that
$$z_i(t) \ge \frac{1}{N^t}, \qquad \forall i = 1, \dots, N, \, \forall t \in \mathbb{N}.$$
3. Exercise 6.1 shows that all entries of $W[t, t+N-1)$ are lower bounded by $1/N^{N-1}$. Use this fact to show that
$$z_i(t) \ge \frac{1}{N^{N-2}}, \qquad \forall i = 1, \dots, N$$
whenever $t \ge N - 1$.
(Hint: Note that $z(t) = W[t-N+1, t)z(t-N+1)$ when $t \ge N-1$, and use $\mathbf{1}^\mathsf{T} z(t-N+1) = \mathbf{1}$.)

4. Conclude that $z_i(t) \geq 1/N^{N-2}$ for all $i = 1, \ldots, N$ and $t \in \mathbb{N}$. $\square$

Theorem 6.1 requires each weight matrix to be row-stochastic. However, the weight matrices in the push-sum method are column-stochastic. In order to exploit Theorem 6.1, we introduce the matrices

$$R(t) = Z(t+1)^{-1} W(t) Z(t), \qquad \forall t \in \mathbb{N},$$

and denote $R[s,t) = R(t-1) \cdots R(s+1) R(s)$. Note that

$$\begin{aligned} R[s,t) &= Z(t)^{-1} W(t-1) Z(t-1) \cdot Z(t-1)^{-1} W(t-2) Z(t-2) \cdots Z(s+1)^{-1} W(s) Z(s) \\ &= Z(t)^{-1} W[s,t) Z(s), \end{aligned}$$

and thus

$$\begin{aligned} \mathsf{X}(t) &= Z(t)^{-1} \mathsf{Y}(t) = Z(t)^{-1} W[0,t) Z(0) Z(0)^{-1} \mathsf{Y}(0) \\ &= R[0,t) \mathsf{Y}(0), \end{aligned}$$

where we used $Z(0) = \mathrm{diag}(\mathbf{1}) = I$. Therefore, the convergence of $\mathsf{X}(t)$ can be directly implied by the convergence of $R[0,t)$ as $t \to \infty$. Also, it is straightforward to check that all entries of $R(t)$ are nonnegative, and $R(t)$ is compatible with the topology of $\mathcal{G}(t)$. Furthermore,

$$R(t)\mathbf{1} = Z(t+1)^{-1} W(t) Z(t) \mathbf{1} = Z(t+1)^{-1} W(t) z(t) = Z(t+1)^{-1} z(t+1) = \mathbf{1},$$

showing that each $R(t)$ is row-stochastic, and

$$R_{ij}(t) = \frac{z_j(t)}{z_i(t+1)} W_{ij}(t) \geq \frac{1/N^{N-2}}{N} \cdot W_{ij}(t),$$

from which we see that

$$R_{ij}(t) > 0 \implies W_{ij}(t) > 0 \implies W_{ij}(t) > \frac{1}{N} \implies R_{ij}(t) \geq \frac{1/N^{N-2}}{N} \cdot \frac{1}{N} = \frac{1}{N^N}.$$

Therefore, the sequence of matrices $(R(t))_{t \in \mathbb{N}}$ satisfies the conditions of Theorem 6.1. By applying Theorem 6.1, we see that there exist $C > 0$, $\sigma \in (0,1)$, $\eta > 0$ and a sequence of vectors $w(s) \in \mathbb{R}^N$, $s = 0, 1, 2, \ldots$ satisfying $w_i(s) > 0$, $\forall i$ and $\mathbf{1}^\mathsf{T} w(s) = 1$ such that

$$\left| \big(R[s,t)\big)_{ij} - w_j(s) \right| = \left| \big(R[s,t) - \mathbf{1} w(s)^\mathsf{T}\big)_{ij} \right| \leq C\sigma^t.$$

As a result, we have

$$\begin{aligned} \left\| \mathsf{X}(t) - \mathbf{1} w(0)^\mathsf{T} \mathsf{Y}(0) \right\|_F &= \left\| \big(R[0,t) - 1 w(0)^\mathsf{T}\big) \mathsf{Y}(0) \right\|_F \\ &= \left\| \big(R[0,t) - 1 w(0)^\mathsf{T}\big) \big(I - 1 w(0)^\mathsf{T}\big) \mathsf{Y}(0) \right\|_F \\ &\leq \left\| R[0,t) - 1 w(0)^\mathsf{T} \right\|_2 \left\| \big(I - 1 w(0)^\mathsf{T}\big) \mathsf{Y}(0) \right\|_F \\ &\leq N \cdot \max_{i,j} \left| \big(R[0,t) - 1 w(0)^\mathsf{T}\big)_{ij} \right| \cdot \left\| \mathsf{Y}(0) - 1 w(0)^\mathsf{T} \mathsf{Y}(0) \right\|_F \\ &\leq CN\sigma^t \cdot \left\| \mathsf{Y}(0) - 1 w(0)^\mathsf{T} \mathsf{Y}(0) \right\|_F, \end{aligned}$$

where in the second step we used $R[0,t) - 1w(0)^\mathsf{T} = \big(R[0,t) - 1w(0)^\mathsf{T}\big)\big(I - 1w(0)^\mathsf{T}\big)$, and in the fourth step we used $\|A\|_2 \le N \max_{i,j} |A_{ij}|$ for any $A \in \mathbb{R}^{N\times N}$.

Our last step in the convergence analysis is to find $w(0)$, which can be done by noting that the vectors $\frac{1}{N}z(s)$ satisfy

$$\begin{aligned}\frac{1}{N}z(t+1)^\mathsf{T}R[s,t) &= \frac{1}{N}z(t+1)^\mathsf{T}Z(t+1)^{-1}W[s,t)Z(s)\\ &= \frac{1}{N}\mathbf{1}^\mathsf{T}W[s,t)Z(s)\\ &= \frac{1}{N}\mathbf{1}^\mathsf{T}Z(s)\\ &= \frac{1}{N}z(s)^\mathsf{T}\end{aligned}$$

for any $t > s \ge 0$. Then since each entry of $z(s)$ is positive and $\mathbf{1}^\mathsf{T}\left(\frac{1}{N}z(s)\right) = 1$, we can apply the results of Exercise 6.3 to conclude that

$$w(s) = \frac{1}{N}z(s), \qquad \forall s \in \mathbb{N},$$

and especially, $w(0) = \frac{1}{N}z(0) = \frac{1}{N}\mathbf{1}$. We finally get

$$\left\|\mathsf{X}(t) - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\mathsf{Y}(0)\right\|_F \le CN\sigma^t \left\|\mathsf{Y}(0) - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\mathsf{Y}(0)\right\|_F,$$

showing that each $x_i(t)$ converges to $\frac{1}{N}\sum_{j=1}^N x_j$ exponentially fast as $t \to \infty$.

Exercise 6.6. For each $t \in \mathbb{N}$, let $W(t)$ be a column-stochastic matrix whose associated directed graph is strongly connected. Suppose all diagonal entries of $W(t)$ are positive, and all positive entries of $W(t)$ are uniformly lower bounded by some constant $\epsilon > 0$. Let

$$z(t) = W[0,t)\mathbf{1}, \quad Z(t) = \operatorname{diag}(z_1(t), \dots, z_N(t)),$$

and

$$R(t) = Z(t+1)^{-1}W(t)Z(t).$$

Show that for arbitrary $b \in \mathbb{R}^N$, $t > s \ge 0$ and $a = R[s,t)b$, we have

$$\left\|a - \frac{1}{N}\mathbf{1}z(t)^\mathsf{T}a\right\| \le CN\sigma^{t-s}\left\|b - \frac{1}{N}\mathbf{1}z(s)^\mathsf{T}b\right\|. \qquad \square$$

6.3 Relaxing the Strong Connectivity Condition

Theorem 6.1 requires that at each time t, the graph $\mathcal{G}(t)$ is strongly connected, which may not be satisfied in some practical scenarios. In this subsection, we present one approach that can relax this condition.

Definition 6.2. A sequence of directed graphs $\mathcal{G}(t) = (\{1,\dots,N\}, \mathcal{E}(t))$, $t \in \mathbb{N}$ is called B-strongly connected, if for any $t \in \mathbb{N}$, the union graph

$$\bigcup_{k=t}^{t+B-1} \mathcal{G}(k) \coloneqq (\{1,\dots,N\}, \mathcal{E}(t) \cup E(t+1) \cup \cdots \cup E(t+B-1))$$

is strongly connected.

The following lemma allows us to relax the strong connectivity condition in Theorem 6.1:

Lemma 6.2. *For each $t \in \mathbb{N}$, let $W(t) \in \mathbb{R}^{N\times N}$ be a matrix with nonnegative entries and let $\mathcal{G}(t) = (\{1,\dots,N\}, \mathcal{E}(t))$ be its associated directed graph. Suppose the following conditions are satisfied:*

1. *$W_{ii} > 0$ for all $i = 1,\dots,N$ and all $t \in \mathbb{N}$.*
2. *There exists $\epsilon > 0$ such that for all $t \in \mathbb{N}$, we have $W_{ij}(t) \geq \epsilon$ whenever $W_{ij}(t) > 0$.*

Then for each $t \in \mathbb{N}$ and $\delta \in \mathbb{N}\backslash\{0\}$, we have

1. *$\big(W[t,t+\delta)\big)_{ij} \geq \epsilon^\delta$ whenever $(j,i) \in \mathcal{E}(t) \cup \cdots \cup \mathcal{E}(t+\delta-1)$.*
2. *$\big(W[t,t+\delta)\big)_{ij} \geq \epsilon^\delta$ whenever $\big(W[t,t+\delta)\big)_{ij} \neq 0$.*

Proof. It's straightforward to see that all entries of $W[t,t+\delta)$ are nonnegative. We shall prove the statement by mathematical induction. The initial case $\delta = 1$ is obvious. Then suppose the statement hold for some $\delta \geq 1$. For the diagonal elements of $W[t,t+\delta+1)$, we have

$$\begin{aligned}\big(W[t,t+\delta+1)\big)_{ii} &= \sum_{k=1}^{N} \big(W(t+\delta)\big)_{ik}\big(W[t,t+\delta)\big)_{ki} \\ &\geq \big(W(t+\delta)\big)_{ii}\big(W[t,t+\delta)\big)_{ii} \geq \epsilon\cdot\epsilon^\delta = \epsilon^{\delta+1}.\end{aligned}$$

For the off-diagonal elements, we have

$$\big(W[t,t+\delta+1)\big)_{ij} = \sum_{k=1}^{N} \big(W(t+\delta)\big)_{ik}\big(W[t,t+\delta)\big)_{kj}. \tag{6.2}$$

If $(j,i) \in \mathcal{E}(t+\delta)$, then

$$\big(W[t,t+\delta+1)\big)_{ij} \geq \big(W(t+\delta)\big)_{ij}\big(W[t,t+\delta)\big)_{jj} \geq \epsilon\cdot\epsilon^\delta = \epsilon^{\delta+1},$$

while if $(j,i) \in \mathcal{E}(t) \cup \cdots \mathcal{E}(t+\delta-1)$, by the induction hypothesis, we get

$$\big(W[t,t+\delta+1)\big)_{ij} \geq \big(W(t+\delta)\big)_{ii}\big(W[t,t+\delta)\big)_{ij} \geq \epsilon\cdot\epsilon^\delta = \epsilon^{\delta+1}.$$

Therefore $\big(W[t,t+\delta+1)\big)_{ij} \geq \epsilon^\delta$ whenever $(j,i) \in \mathcal{E}(t)\cup\cdots\cup\mathcal{E}(t+\delta-1)\cup\mathcal{E}(t+\delta)$. Now suppose $\big(W[t,t+\delta+1)\big)_{ij} \neq 0$. By (6.2) and the nonnegativity of the entries, there must exist some index k^* such that $\big(W(t+\delta)\big)_{ik^*} > 0$ and $\big(W[t,t+\delta)\big)_{k^*j} > 0$. We then have $\big(W(t+\delta)\big)_{ik^*} > \epsilon$ by our assumption on $W(t+\delta)$ and $\big(W[t,t+\delta)\big)_{k^*j} \geq \epsilon^\delta$ by the induction hypothesis. As a result,

$$\big(W[t,t+\delta+1)\big)_{ij} \geq \big(W(t+\delta)\big)_{ik^*}\big(W[t,t+\delta)\big)_{k^*j} \geq \epsilon\cdot\epsilon^\delta = \epsilon^{\delta+1}.$$

We can now complete the proof by induction on δ. □

As a corollary, we have the following extension of Theorem 6.1:

Theorem 6.2. *For each $t \in \mathbb{N}$, let $W(t) \in \mathbb{R}^{N\times N}$ be a weight matrix, and let $\mathcal{G}(t) = (\{1,\dots,N\}, \mathcal{E}(t))$ be its associated directed graph. Suppose the following conditions are satisfied:*

1. *The sequence of graphs $(\mathcal{G}(t))_{t\in\mathbb{N}}$ is B-strongly connected for some positive integer B.*
2. *Each $W(t)$ is row-stochastic.*
3. *$W_{ii}(t) > 0$ for all $i = 1,\dots,N$ and all $t \in \mathbb{N}$.*
4. *There exists $\epsilon > 0$ such that for any $t \in \mathbb{N}$, $W_{ij}(t) \geq \epsilon$ whenever $W_{ij} > 0$.*

Then there exist $C > 0$, $\sigma \in (0,1)$, $\eta > 0$ and a sequence of vectors

$$w(s) \in \left\{w \in \mathbb{R}^N : w_i \geq \eta,\ \forall i \text{ and } \mathbf{1}^\mathsf{T} w = 1\right\},\ s = 0,1,2,\dots$$

such that

$$\left|\big(W[s,t)\big)_{ij} - w_j(s)\right| \leq C\sigma^{t-s}$$

for any i,j and any $t > s \geq 0$. The constants C, σ and η only depend on the number of agents N, the lower bound of the positive entries ϵ and B. Furthermore, $w(s) = \frac{1}{N}\mathbf{1}$ if each $W(t)$ is doubly-stochastic.

Proof. Fix $s \in \mathbb{N}$. For each $k \in \mathbb{N}$, denote

$$\begin{aligned} D(k;s) &= W(s+(k+1)B-1)\cdots W(s+kB+1)\cdot W(s+kB) \\ &= W[s+kB, s+(k+1)B), \end{aligned}$$

and let $\tilde{\mathcal{G}}(k;s) = \{\{1,\dots,N\}, \tilde{\mathcal{E}}(k,s)\}$ be the directed graph satisfying

$$(j,i) \in \tilde{\mathcal{E}}(k;s) \iff D_{ij}(k;s) > 0.$$

By Lemma 6.2, $D_{ij}(k;s) \geq \epsilon^B$ whenever $D_{ij}(k;s) > 0$, and

$$\bigcup_{i=0}^{B-1} \mathcal{E}(s+kB+i) \subseteq \tilde{\mathcal{E}}(k;s).$$

Since $(\mathcal{G}(t))_{t\in\mathbb{N}}$ is B-strongly connected, each $\tilde{\mathcal{G}}(k;s)$ is strongly connected. Moreover, it's straightforward to check that $D(k;s)\mathbf{1} = \mathbf{1}$ for all $k \in \mathbb{N}$, and that every node in $\tilde{\mathcal{G}}(k;s)$ has a self-loop. By Theorem 6.1, there exist $\tilde{C}_s > 0$, $\tilde{\sigma}_s \in (0,1)$, $\eta_s > 0$ and a vector $w(s) \in \mathbb{R}^N$ satisfying

$$w_j(s) \geq \eta_s,\ \forall j \text{ and } \mathbf{1}^\mathsf{T} w(s) = 1,$$

such that

$$\left|\big(D(k-1;s)D(k-2;s)\cdots D(0;s)\big)_{ij} - \tilde{w}_j(s)\right| \leq \tilde{C}_s\tilde{\sigma}_s^k$$

for any i,j and any $k > 0$. Since $\tilde{C}_s$, $\tilde{\sigma}_s$ and η_s only depends on N and ϵ^B, we can drop their subscripts s.

Now let $t > s \geq 0$ be arbitrary, and let $\tau = \lfloor (t-s)/B \rfloor$. When $\tau > 0$, we have

$$W[s,t] = W[s+\tau B, t] D(\tau - 1; s) \cdots D(0; s),$$

(we let $W[s+\tau_t, t] = I$ if $s + \tau_t = t$). Note that

$$\begin{aligned}
\big(W[s,t]\big)_{ij} &= \sum_{k=1}^{N} \big(W[s+\tau B, t]\big)_{ik} \big(D(\tau-1;s)\cdots D(0;s)\big)_{kj} \\
&\leq \max_{1\leq k\leq N} \big(D(\tau-1;s)\cdots D(0;s)\big)_{kj} \cdot \sum_{k=1}^{N} \big(W[s+\tau B,t]\big)_{ik} \\
&= \max_{1\leq k\leq N} \big(D(\tau-1;s)\cdots D(0;s)\big)_{kj},
\end{aligned}$$

where we used $W[s+\tau B, t]\mathbf{1} = \mathbf{1}$, and similarly

$$\begin{aligned}
\big(W[s,t]\big)_{ij} &\leq \min_{1\leq k\leq N} \big(D(\tau-1;s)\cdots D(0;s)\big)_{kj} \cdot \sum_{k=1}^{N} \big(W[s+\tau B,t]\big)_{ik} \\
&= \min_{1\leq k\leq N} \big(D(\tau-1;s)\cdots D(0;s)\big)_{kj}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\Big|\big(W[s,t]\big)_{ij} - w_j(s)\Big| \leq\ & \max\Bigg\{ \Big| \max_{1\leq k\leq N} \big(D(\tau-1;s)\cdots D(0;s)\big)_{kj} - w_j(s) \Big|, \\
& \qquad \Big| \min_{1\leq k\leq N} \big(D(\tau-1;s)\cdots D(0;s)\big)_{kj} - w_j(s) \Big| \Bigg\} \\
\leq\ & \tilde{C}\tilde{\sigma}^{\tau} \leq \tilde{C}\tilde{\sigma}^{(t-s)/B}.
\end{aligned}$$

When $\tau = 0$, we have $t - s \leq B - 1$ and

$$\begin{aligned}
\Big|\big(W[s,t]\big)_{ij} - w_j(s)\Big| &\leq \Big|\big(W[s,t]\big)_{ij}\Big| + |w_j(s)| \leq 2 \\
&\leq \frac{2}{\tilde{\sigma}^{1-1/B}} \tilde{\sigma}^{(t-s)/B}
\end{aligned}$$

By letting $C = \max\{\tilde{C}, 2/\tilde{\sigma}^{1-1/B}\}$, $\sigma = \tilde{\sigma}^{1/B}$, we get the desired results. □

By using Theorem 6.2, we can prove that for the push-sum algorithm (6.1), $x_i(t) - \frac{1}{N}\sum_{j=1}^{N} x_j$ converges exponentially to 0 when the sequence of graphs $(\mathcal{G}_t)_{t\in\mathbb{N}}$ is only B-strongly connected for some positive integer B.

6.4 Distributed Optimization over Time-Varying Communication Networks

In this section, we present a distributed algorithm for consensus optimization over time-varying networks from the work [Saadatniaki et al., 2020].

Consider the consensus optimization problem

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

where each local cost function f_i is continuously differentiable. We assume that an optimal solution to the above problem exists and denote it by $x^* \in \mathbb{R}^d$. The N agents are connected by a time-varying communication network whose topology at time t is described by the directed graph $\mathcal{G}(t) = (\{1,\dots,N\}, \mathcal{E}(t))$.

The algorithm proposed in [Saadatniaki et al., 2020], named the TV-$\mathcal{AB}$ algorithm in the paper, employs two weight matrices $A(t)$ and $B(t)$ compatible with the communication network at each time step t, with $A(t)$ being row-stochastic and $B(t)$ being column stochastic. The iterations are given by

$$\begin{aligned} x_i(t+1) &= \sum_{j=1}^{N} A_{ij}(t)x_j(t) - \eta g_i(t), \\ g_i(t+1) &= \sum_{j=1}^{N} B_{ij}(t)g_j(t) + \nabla f_i(x_i(t+1)) - \nabla f_i(x_i(t)), \end{aligned} \tag{6.3}$$

or, written compactly,

$$\begin{aligned} \mathsf{X}(t+1) &= A(t)\mathsf{X}(t) - \eta\mathsf{G}(t), \\ \mathsf{G}(t+1) &= B(t)\mathsf{G}(t) + \nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t)). \end{aligned}$$

The algorithm (6.3) is an extension of the following push-pull gradient algorithm [Pu et al., 2021] to time-varying communication networks:

$$\begin{aligned} \mathsf{X}(t+1) &= A\mathsf{X}(t) - \eta\mathsf{G}(t), \\ \mathsf{G}(t+1) &= B\mathsf{G}(t) + \nabla F(\mathsf{X}(t+1)) - \nabla F(\mathsf{X}(t)), \end{aligned} \tag{6.4}$$

which is an extension of the basic gradient tracking algorithm (4.1) to (time-invariant) directed communication networks.

Remark 6.1. Note that in both TV-$\mathcal{AB}$ (6.3) and the push-pull gradient algorithm (6.4), there are no extra variables $z_i(t)$ that are used to scale the entries of the variables $x_i(t)$. An intuitive explanation is as follows: First consider the simpler case with time-invariant strongly connected networks. As t becomes sufficiently large, by the Perron–Frobenius Theorem, we have

$$B^t \approx \nu\mathbf{1}^{\mathsf{T}},$$

where $\nu \in \mathbb{R}^N$ is a vector with positive entries that sum up to 1. We can then intuitively expect that

$$\mathsf{G}(t) \approx \nu\mathbf{1}^{\mathsf{T}}\nabla F(\mathsf{X}(t))$$

when t is sufficiently large. We also have

$$A^t \approx \mathbf{1}\mu^{\mathsf{T}}$$

for sufficiently large t, where $\mu \in \mathbb{R}^N$ is a vector with positive entries that sum up to 1. As a result, we intuitively expect that, for sufficiently large t,

$$\mathsf{X}(t) \approx \mathbf{1}\mu^\mathsf{T}\mathsf{X}(t).$$

Then the iteration of $\mathsf{X}(t)$ gives

$$\mu^\mathsf{T}\mathsf{X}(t+1) \approx \mu^\mathsf{T}\mathsf{X}(t) - \eta\mu^\mathsf{T}\nu\mathbf{1}^\mathsf{T}\nabla F(\mathbf{1}\mu^\mathsf{T}\mathsf{X}(t)),$$

and if we denote $\tilde{x}(t) = \mu^\mathsf{T}\mathsf{X}(t)$ and $\tilde{\eta} = N\eta\mu^\mathsf{T}\nu$, we get

$$\tilde{x}(t+1) \approx \tilde{x}(t) - \tilde{\eta}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\tilde{x}(t)),$$

which just follows an approximate gradient descent update. For the time-varying setting, the vectors μ and ν will become dependent on t, and the analysis will rely on Theorem (6.1) or (6.2) to bound the consensus errors and the optimality gap.

The following theorem establishes the convergence of TV-$\mathcal{AB}$ for strongly convex problems:

Theorem 6.3 ([Saadatniaki et al., 2020]). *Suppose each f_i is L-smooth, and the global objective function f is μ-strongly convex. Suppose the weight matrices $A(t)$ and $B(t)$ are compatible with the graph $\mathcal{G}(t)$, and further the following conditions hold:*

1. *Each $A(t)$ is row-stochastic, and each $B(t)$ is column-stochastic, i.e., all entries of $A(t)$ and $B(t)$ are nonnegative, and $A(t)\mathbf{1} = B(t)^\mathsf{T}\mathbf{1} = \mathbf{1}$.*
2. *All nodes of $\mathcal{G}(t)$ have self-loops, i.e., $(i,i) \in \mathcal{E}(t)$ for all i and all t.*
3. *The sequence of graphs $(\mathcal{G}(t))_{t\in\mathbb{N}}$ is B-strongly connected for some positive integer B.*
4. *There exists $\epsilon > 0$ such that $A_{ij}(t) \geq \epsilon$ and $B_{ij}(t) \geq \epsilon$ whenever $(j,i) \in \mathcal{E}(t)$.*

Then we have

$$\|\mathsf{X}(t) - \mathbf{1}x^*\|_F \leq O(\rho(\eta)^t),$$

where $\rho(\eta) > 0$ will be strictly less than 1 for sufficiently small η.

Notes on References

The materials in this chapter are based on a series of papers that study distributed averaging/consensus optimization methods over time-varying communication networks, including the early seminal work [Tsitsiklis et al., 1986] that established a large proportion of the theoretical tools for analyzing time-varying networks, and also the more recent works including [Jadbabaie et al., 2003], [Nedić and Ozdaglar, 2009], [Nedić et al., 2009], [Nedić and Olshevsky, 2014], [Nedić and Liu, 2017], [Nedić et al., 2017], [Saadatniaki et al., 2020], etc. Some of the technical details, including Exercise 6.1 and Part II of the proof of Theorem 6.1, are adapted from or inspired by [Bullo, 2022, Chapter 12].

6.A Proof of Theorem 6.1

Part I: Row-stochastic case. We first consider the case where each $W(t)$ is not assumed be doubly stochastic. We fix an arbitrary $s \in \mathbb{N}$, and denote

$$\begin{aligned} D(k;s) &= W(s+(k+1)(N-1)-1)\cdots W(s+k(N-1)+1)\cdot W(s+k(N-1)) \\ &= W[s+k(N-1), s+(k+1)(N-1)), \qquad k \in \mathbb{N}. \end{aligned}$$

By the results of Exercise 6.1, all entries of $D(k,s)$ are greater than or equal to ϵ^{N-1}.

Now let $u(0) \in \mathbb{R}^N$ be arbitrary, and let

$$u(k+1) = D(k;s)u(k)$$

for each $k \geq 0$. We denote $h_{\min}(k) = \arg\min_h z_h(k)$ and $h_{\max}(k) = \arg\max_h z_h(k)$ for any $k \geq 0$ (ties are broken arbitrarily). Then

$$\begin{aligned} u_i(k+1) - u_j(k+1) &= \sum_{h=1}^{N} (D_{ih}(k;s) - D_{jh}(k;s))u_h(k) \\ &= \sum_{h=1}^{N} (D_{ih}(k;s) - D_{jh}(k;s)) \left(u_h(k) - u_{h_{\min}(k)}(k)\right) \\ &= \sum_{h \neq h_{\min}(k)} (D_{ih}(k;s) - D_{jh}(k;s)) \left(u_h(k) - u_{h_{\min}(k)}(k)\right) \\ &\leq \left(u_{h_{\max}(k)}(k) - u_{h_{\min}(k)}(k)\right) \sum_{h \neq h_{\min}(k)} (D_{ih}(k;s) - D_{jh}(k;s)) \\ &\leq \left(u_{h_{\max}(k)}(k) - u_{h_{\min}(k)}(k)\right) \sum_{h \neq h_{\min}(k)} D_{ih}(k;s) \\ &= (u_{h_{\max}(k)}(k) - u_{h_{\min}(k)}(k))(1-\epsilon^{N-1}), \end{aligned}$$

where in the second step we used $\sum_{h=1}^{N} D_{ih}(k,s) = \sum_{h=1}^{N} D_{jh}(k,s) = 1$. As a result,

$$\max_{1\leq h\leq N} u_h(k+1) - \min_{1\leq h\leq N} u_h(k+1) \leq (1-\epsilon^{N-1}) \left(\max_{1\leq h\leq N} u_h(k) - \min_{1\leq h\leq N} u_h(k)\right)$$

and consequently

$$\max_{1\leq h\leq N} u_h(k) - \min_{1\leq h\leq N} u_h(k) \leq (1-\epsilon^{N-1})^k \left(\max_{1\leq h\leq N} u_h(0) - \min_{1\leq h\leq N} u_h(0)\right).$$

Now, to show that $u(k)$ converges as $k \to \infty$, we note that for all i,

$$\begin{aligned} u_i(k+1) &= \sum_{j=1}^{N} D_{ij}(k;s)u_j(k) \\ &= \sum_{j=1}^{N} D_{ij}(k;s) \left(u_j(k) - u_{h_{\min}(k)}(k)\right) + u_{h_{\min}(k)}(k) \sum_{j=1}^{N} D_{ij}(k;s) \end{aligned}$$

$$\leq \left(u_{h_{\max(k)}}(k) - u_{h_{\min(k)}}(k)\right) \sum_{j=1}^{N} D_{ij}(k;s) + u_{h_{\min(k)}}(k)$$
$$= u_{h_{\min(k)}}(k) + \left(u_{h_{\max(k)}}(k) - u_{h_{\min(k)}}(k)\right),$$

which implies

$$u_{h_{\min(k+1)}}(k+1) - u_{h_{\min(k)}}(k) \leq (1-\epsilon^{N-1})^k \left(\max_{1\leq h\leq N} u_h(0) - \min_{1\leq h\leq N} u_h(0)\right).$$

Furthermore,

$$u_{h_{\min(k+1)}}(k+1) = \sum_{j=1}^{N} D_{h_{\min(k+1)},j}(k;s) u_j(k)$$
$$\geq u_{h_{\min(k)}}(k) \sum_{j=1}^{N} D_{h_{\min(k+1)},j}(k;s) = u_{h_{\min(k)}}(k).$$

Therefore

$$0 \leq u_{h_{\min(k+1)}}(k+1) - u_{h_{\min(k)}}(k) \leq (1-\epsilon^{N-1})^k \left(\max_{1\leq h\leq N} u_h(0) - \min_{1\leq h\leq N} u_h(0)\right),$$

which further implies

$$0 \leq u_{h_{\min(k+j)}}(k+j) - u_{h_{\min(k)}}(k) \leq \sum_{i=0}^{j-1} (1-\epsilon^{N-1})^{k+i} \left(\max_{1\leq h\leq N} u_h(0) - \min_{1\leq h\leq N} u_h(0)\right)$$
$$\leq \frac{(1-\epsilon^{N-1})^k}{\epsilon^{N-1}} \left(\max_{1\leq h\leq N} u_h(0) - \min_{1\leq h\leq N} u_h(0)\right)$$

for any $k, j \in \mathbb{N}$. Therefore $(u_{h_{\min(k)}}(k))_{k\in\mathbb{N}}$ is a Cauchy sequence and converges to some $u(\infty)$ as $k \to \infty$. By letting $j \to \infty$ in the above inequality, we get

$$|u_{h_{\min(k)}}(k) - u(\infty)| \leq \frac{(1-\epsilon^{N-1})^k}{\epsilon^{N-1}} \left(\max_{1\leq h\leq N} u_h(0) - \min_{1\leq h\leq N} u_h(0)\right).$$

Then for any i, we have

$$|u_i(k+1) - u(\infty)| \leq \left|u_{h_{\max(k)}}(k) - u_{h_{\min(k)}}(k)\right| + \left|u_{h_{\min(k)}}(k) - u(\infty)\right|$$
$$\leq \left(1+\epsilon^{1-N}\right)(1-\epsilon^{N-1})^k \left(\max_{1\leq h\leq N} u_h(0) - \min_{1\leq h\leq N} u_h(0)\right).$$

Now by letting $u(0) = e_j$ for each standard basis vector $e_j \in \mathbb{R}^N$ and let the corresponding $u(\infty)$ be denoted by $w_j(s)$, we get

$$\left|\left(D(k-1;s)\cdots D(0;s)\right)_{ij} - w_j(s)\right| \leq \left(1+\epsilon^{1-N}\right)(1-\epsilon^{N-1})^k.$$

Now let $t > s \geq 0$ be arbitrary. We can now use similar techniques in the second half of the proof of Theorem 6.2 to prove that

$$\left|\left(W[s,t)\right)_{ij} - w_j(s)\right| \leq C(1-\epsilon^{N-1})^{t-s}$$

for some $C > 0$ that depends only on ϵ and N, and we omit the details here. To show that $w_j(s)$ has a uniform lower bound η, we use mathematical induction to show that $\big(D(k;s)\cdots D(0;s)\big)_{ij} \geq \epsilon^{N-1}$ for all $k \in \mathbb{N}$. The initial case can be justified by Exercise 6.1. Now suppose

$$\big(D(k;s)\cdots D(0;s)\big)_{ij} \geq \epsilon^{N-1}$$

for some $k \in \mathbb{N}$. We have

$$\begin{aligned}\big(D(k+1;s)\cdots D(0;s)\big)_{ij} &= \sum_{h=1}^{N} D_{ih}(k+1;s)\big(D(k;s)\cdots D(0;s)\big)_{hj} \\ &\geq \sum_{h=1}^{N} D_{ih}(k+1;s)\cdot\epsilon^{N-1} = \epsilon^{N-1}.\end{aligned}$$

By mathematical induction we get $\big(D(k;s)\cdots D(0;s)\big)_{ij} \geq \epsilon^{N-1}$ for all $k \in \mathbb{N}$, and by taking the limit $k \to \infty$ we get $w_j(s) \geq \epsilon^{N-1}$.

Part II. Doubly stochastic case. We now impose the further assumption that each $W(t)$ is doubly stochastic. For any $\mathcal{E} \subseteq \{1,\dots,N\}\times\{1,\dots,N\}$, denote

$$\mathcal{P}(\mathcal{E}) = \{P \in \mathbb{R}^{N\times N} : P = P^{\mathsf{T}},\, P\mathbf{1} = \mathbf{1},\, P_{ij} \geq \epsilon^2,\, \forall (j,i) \in \mathcal{E},\, P_{ij} = 0,\, \forall (j,i) \notin \mathcal{E}\},$$

and let $\mathcal{C}$ denote the subset of the power set of $\{1,\dots,N\}\times\{1,\dots,N\}$ such that $\mathcal{E} \in \mathcal{C}$ if and only if $(\{1,\dots,N\},\mathcal{E})$ is undirected, strongly connected and $(i,i) \in \mathcal{E}$ for all $i = 1,\dots,N$. It's not hard to see that $\mathcal{C}$ is a finite set, and consequently

$$\mathcal{P}_{\mathcal{C}} \coloneqq \bigcup_{\mathcal{E}\in\mathcal{C}} \mathcal{P}(\mathcal{E})$$

is a compact set. Moreover, for any $P \in \mathcal{P}_{\mathcal{C}}$, by the Perron-Frobenius Theorem, we have

$$\left\| P - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}} \right\|_2 < 1.$$

Therefore

$$\overline{\sigma} \coloneqq \sup_{P\in\mathcal{P}_{\mathcal{C}}} \left\| P - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}} \right\|_2 < 1.$$

It's not hard to see that $\overline{\sigma}$ only depends on ϵ and N.

Now let $u(0) \in \mathbb{R}^N$ be arbitrary, and let

$$u(t+1) = W(t)u(t), \qquad t \in \mathbb{N}.$$

Since each $W(t)$ is doubly stochastic, we have

$$\frac{1}{N}\mathbf{1}^{\mathsf{T}}u(t+1) = \frac{1}{N}\mathbf{1}^{\mathsf{T}}W(t)u(t) = \frac{1}{N}\mathbf{1}^{\mathsf{T}}u(t) = \cdots = \frac{1}{N}\mathbf{1}^{\mathsf{T}}u(0).$$

Define

$$V(t) \coloneqq \frac{1}{2}\left\| u(t) - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}}u(0) \right\|^2 = \frac{1}{2}\left\| \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^{\mathsf{T}} \right) u(t) \right\|^2.$$

We then have

$$\begin{aligned} V(t+1) &= \frac{1}{2}\left\|\left(I-\frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)W(t)u(t)\right\|^2 \\ &= \frac{1}{2}\left\|\left(W(t)-\frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)\left(I-\frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)u(t)\right\|^2 \\ &\leq \frac{1}{2}\left\|W(t)-\frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2^2 \cdot \left\|\left(I-\frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right)u(t)\right\|_2^2 \\ &= \frac{1}{2}\left\|W(t)^\mathsf{T}W(t)-\frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}\right\|_2 V(t). \end{aligned}$$

It's not hard to check that $W(t)^\mathsf{T}W(t) \in \mathcal{P}_\mathcal{C}$, and therefore

$$V(t+1) \leq \overline{\sigma} V(t).$$

We can now see that $\|u(t) - \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}u(0)\|$ converges exponentially to zero with rate $\sqrt{\overline{\sigma}}$. By the arbitrariness of $u(0) \in \mathbb{R}^N$, we conclude that

$$\max_{i,j}\left|\big(W[0,t)\big)_{ij} - \frac{1}{N}\right| \leq C\overline{\sigma}^{t/2}$$

for some constant $C > 0$ that only depends on N and ϵ. The bound for $\left|\big(W[s,t)\big)_{ij} - \frac{1}{N}\right|$ for arbitrary $t > s \geq 0$ is straightforward.

Remark 6.2. It can be shown that the rate $\overline{\sigma}$ has an explicit upper bound:

$$\overline{\sigma} \leq 1 - \frac{\epsilon}{2N^2}.$$

We refer to [Nedić et al., 2009, Lemma 9] for more details.

Bibliography

[Bullo, 2022] Bullo, F. (2022). *Lectures on Network Systems*. Kindle Direct Publishing, 1.6 edition.

[Jadbabaie et al., 2003] Jadbabaie, A., Lin, J., and Morse, A. S. (2003). Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on automatic control*, 48(6):988–1001.

[Nedić and Liu, 2017] Nedić, A. and Liu, J. (2017). On convergence rate of weighted-averaging dynamics for consensus problems. *IEEE Transactions on Automatic Control*, 62(2):766–781.

[Nedić and Olshevsky, 2014] Nedić, A. and Olshevsky, A. (2014). Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615.

[Nedić et al., 2009] Nedić, A., Olshevsky, A., Ozdaglar, A., and Tsitsiklis, J. N. (2009). On distributed averaging algorithms and quantization effects. *IEEE Transactions on automatic control*, 54(11):2506–2517.

[Nedić et al., 2017] Nedić, A., Olshevsky, A., and Shi, W. (2017). Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633.

[Nedić and Ozdaglar, 2009] Nedić, A. and Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61.

[Pu et al., 2021] Pu, S., Shi, W., Xu, J., and Nedić, A. (2021). Push–pull gradient methods for distributed optimization in networks. *IEEE Transactions on Automatic Control*, 66(1):1–16.

[Saadatniaki et al., 2020] Saadatniaki, F., Xin, R., and Khan, U. A. (2020). Decentralized optimization over time-varying directed graphs with row and column-stochastic matrices. *IEEE Transactions on Automatic Control*, 65(11):4769–4780.

[Tsitsiklis et al., 1986] Tsitsiklis, J., Bertsekas, D., and Athans, M. (1986). Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812.

Chapter 7

Federated Learning from a Distributed Optimization Viewpoint

7.1 Problem Setup

Federated learning is a relatively new topic in machine learning that has caught much interest across multiple disciplines. In the survey paper [Kairouz et al., 2021], the authors propose the following definition of federated learning:

> ***Federated learning*** *is a machine learning setting where multiple entities (clients) collaborate in solving a machine learning problem, under the coordination of a central server or service provider. Each client's raw data is stored locally and not exchanged or transferred; instead, focused updates intended for immediate aggregation are used to achieve the learning objective.*

Here, by focused updates the authors mean "updates narrowly scoped to contain the minimum information necessary for the specific learning task at hand", and "aggregation is performed as early as possible in the service of data minimization".

In this chapter, we shall further narrow down our investigation by considering only the so-called *horizontal* federated learning setting and viewing it from a distributed optimization perspective. Specifically, let us consider a cross-device learning task where a large number of mobile devices (such as phones, tablets, mobile sensors, etc.) are coordinated by a central server to train a machine learning model, with datasets distributed and locally stored at each mobile device. We model the task as a distributed optimization problem under the server-worker setup. We let N denote the number of workers (or clients, which seems a more common terminology in federated learning) that represent the mobile devices, and suppose each client owns a dataset $\mathcal{D}_i$, $i = 1, \ldots, N$. The communication network will have a star topology with the server placed at

the center node. Note that for this cross-device learning task, the communication links, especially the ones from the devices to the server, will have very limited bandwidth and reliability. The goal is to solve the following optimization problem:

$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{|\mathcal{D}|}\sum_{z\in\mathcal{D}} \ell(z;x), \tag{7.1}$$

where $\ell(z;x)$ is the loss function that quantifies how well the parameterized model x fits the single sample z, and $\mathcal{D}$ is the (disjoint) union of all local datasets $\mathcal{D}_1,\ldots,\mathcal{D}_N$. We let $p_i = |\mathcal{D}_i|/|\mathcal{D}|$ denote the proportion of the local dataset $\mathcal{D}_i$ in the combined dataset $\mathcal{D}$.

Note that the federated learning setting requires that each dataset $\mathcal{D}_i$ be locally stored at client i and cannot be revealed to either the central server or other clients. In order to meet this requirement, we consider reformulating the federated learning problem (7.1) as

$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{N}\sum_{i=1}^N f_i(x), \quad f_i(x) = \frac{p_i N}{|\mathcal{D}_i|}\sum_{z\in\mathcal{D}_i}\ell(z;x) = \mathbb{E}_{z\sim\mathcal{P}_i}[p_i N \ell(z;x)],$$

where $\mathcal{P}_i$ denotes the uniform distribution over the set $\mathcal{D}_i$. This formulation suggests one approach to implement the stochastic gradient descent iteration:

$$x(t+1) = x(t) - \eta_t \cdot \frac{1}{N}\sum_{i=1}^N g_i(t), \tag{7.2}$$

where each $g_i(t)$ is a stochastic gradient of $f_i(x)$, which can be constructed by agent i individually via sampling a minibatch of B independent points $z_{i,b}(t)\sim\mathcal{P}_i$, $b=1,\ldots,B$ and form

$$g_i(t) = \frac{p_i N}{B}\sum_{b=1}^B \nabla_x \ell(z_{i,b}(t);x(t)). \tag{7.3}$$

By further incorporating the communication network, we get the following prototypical algorithm for solving (7.1):

1. **The server** sends the current iterate $x(t)$ to all clients.
2. **Client** i samples a minibatch $(z_{i,b}(t))_{b=1}^B$ with $z_{i,b}(t)\sim\mathcal{P}_i$ and construct $g_i(t)$ by (7.3).
3. **Client** i sends $g_i(t)$ back to the server.
4. **The server** collects all $g_i(t)$ and conducts the SGD update (7.2).
5. Set $t\leftarrow t+1$ and go back to Step 1.

The above approach seems to be reasonable if implemented in a data center where the communication between clients and the server is high-speed and reliable and the number of clients is typically $\lesssim 1000$. However, if the number of clients N becomes much larger (say $N\sim 10^6$ to 10^{10}) and the communication links have limited bandwidth and reliability, which is the case for cross-device federated learning tasks, the above procedure will soon encounter the following bottlenecks:

- The upload of the local gradients $g_i(t)$ to the server may take a very long time, due to both the huge N and the limited bandwidth. Specifically, if each d-dimensional vector $g_i(t)$ is encoded by Q bits, then the number of bits received by the server per SGD update is NQ, and it can be very expensive to handle such a large amount of incoming data. Considering further that the communication links have limited bandwidth, the time it takes for the server to collect all local gradients may easily unacceptable.
- Not all clients may be available for contributing their local gradients throughout the optimization procedure. Some clients may fail or drop out at some time due to unreliable communication, low battery level, etc. A client may also become slow if the device has already been running some computationally intensive apps.

In the following sections, we present a basic approach that is intended to address or alleviate some of the above bottlenecks.

7.2 The Federated Averaging Algorithm

The federated averaging algorithm (abbreviated as FedAvg) [McMahan et al., 2017] is one of the most common approaches to optimization for federated learning. The basic steps of FedAvg are given as follows:

1. **The server** uniformly samples a subset $\mathcal{S}(t) \subseteq \{1, \ldots, N\}$ with size $|\mathcal{S}(t)| = S$.
2. **The server** broadcasts the current iterate $x(t)$ to the clients in the subset $\mathcal{S}(t)$.
3. **for** each $i \in \mathcal{S}(t)$ in parallel **do**
4. **Client** i sets $x_i^t(0) = x(t)$.
5. **for** $k = 0, \ldots, K-1$ **do**
6. **Client** i samples a minibatch $\mathcal{B}_i^t(k) = (z_{i,b}^t(k))_{b=1}^B$ of B i.i.d. samples from $\mathcal{P}_i$.
7. **Client** i constructs

$$g_i^t(k) = \frac{p_i N}{B} \sum_{b=1}^{B} \nabla_x \ell(z_{i,b}^t(k); x_i^t(k)),$$

and updates

$$x_i^t(k+1) = x_i^t(k) - \eta\, g_i^t(k).$$

8. **end for**
9. **Client** $i \in \mathcal{S}(t)$ uploads $\Delta x_i(t+1) = x_i^t(K) - x(t)$ to the server.
10. **end for**
11. **The server** updates the iterate by

$$x(t+1) = x(t) + \frac{\alpha}{S} \sum_{i \in \mathcal{S}(t)} \Delta x_i(t+1).$$

12. Set $t \leftarrow t+1$ and go back to Line 1 unless $t = T-1$.

The parameters of the algorithm include the step sizes $\eta > 0$ for local SGD update and $\alpha > 0$ for aggregation, the number of clients queried per iteration S, the batch size B, the number of inner local SGD updates K, and the total number of outer iterations T.

FedAvg attempts to address the issues caused by large N and limited communication bandwidth by the following approaches:

- For each t, the server only queries a subset of the group of clients with size S. This allows partial participation of the clients in each round of communication and reduces the burden on the communication links between consecutive updates on $x(t)$.
- After Client i samples a minibatch and obtains a local stochastic gradient $g_i(t)$, the client does not immediately upload $g_i(t)$ to the server but rather performs local SGD updates on its own. Only after K iterations of local SGD will Client i upload the vector $x_i(t+1) = \texttt{ClientUpdate}(i, x(t))$ to the server. In this way, the frequency of communication will be reduced.

As a result, assuming that uploading each d-dimensional vector $\Delta x_i(t+1)$ takes Q bits, we can see that the number of bits received by the server per local SGD update is SQ/K, which can be substantially smaller than NQ.

Remark 7.1. Note that in FedAvg, the subset of participating clients $\mathcal{S}(t)$ is randomly selected by the server, and it is assumed that each selected client is able to upload $\Delta x_i(t+1)$ successfully for every t. This assumption may not hold in practical applications, and FedAvg may still encounter issues caused by communication failure or dropped-out devices. However, suppose for each t, the delays between the server broadcasting $x(t)$ and the server receiving $\Delta x_i(t+1)$ for different i can be modeled as i.i.d. random variables taking values in $(0, +\infty]$, where a $+\infty$ delay means upload failure, and the delays across different time steps t are independent. Then in this case, the server may broadcast $x(t)$ to all (or a sufficiently large random subset of) clients, collect the first S arriving $\Delta x_i(t+1)$ and discard all other $\Delta x_i(t+1)$; in the rare case when less than S devices respond, the server collects all $\Delta x_i(t+1)$ that have been received. The resulting set of participating clients will be very close to being uniform sampled from $\{1, \ldots, N\}$.

Remark 7.2. In the original version of FedAvg, the step size for global aggregation α is simply chosen to be 1. Here we allow α to take other positive values, which was proposed by [Karimireddy et al., 2020]. There are also other variants of FedAvg; some adopt different strategies for choosing the subset of participating clients $\mathcal{S}(t)$, and some employ weighted average when aggregating the local updates $\Delta x_i(t+1)$.

In sum, the whole procedure of FedAvg can be summarized by the following framework:

1. Client selection: This step allows partial participation of clients. The original FedAvg employs simple uniform sampling (without replacement), but in practice one may employ more complicated selection strategy to better address the issue of unreliable devices and communication links.
2. Broadcast: The server broadcasts the current model parameter to the selected clients.
3. Client computation: Each selected client computes an update of the model parameter, using its locally stored data.

4. Aggregation: The server collects the clients' updates of the model parameters. Stragglers that cannot upload the updates in time will be dropped.
5. Model update: The server updates the global model parameter based on the aggregated local updates.

This framework can serve as a very good starting point for designing federated learning algorithms, and also provides sufficient flexibility to accommodate many other techniques including privacy preservation measures, gradient compression for efficient communication, etc.

7.3 Convergence of Federated Averaging

It turns out that establishing convergence for FedAvg is not an easy task, especially when the local dataset distributions $\mathcal{P}_i$ are different. In the following, we discuss the convergence of FedAvg for two situations, one with $\mathcal{P}_i$ being the same for all i, and the other with $\mathcal{P}_i$ being different from each other.

Identical Local Distributions

Rigorously speaking, the distributions $\mathcal{P}_i$ will be different as long as the local datasets $\mathcal{D}_i$ are not the same. However, when each local dataset $\mathcal{D}_i$ contains i.i.d. samples generated from the same underlying distribution $\mathcal{P}$ independently (for example, the local data are generated by the same experimental procedures independently under identical environment), and the size $|\mathcal{D}_i|$ is sufficiently large for all i, we may approximate $\mathcal{P}_i$ by $\mathcal{P}$ in the theoretical analysis. In this case, the global objective function becomes

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{z\sim\mathcal{P}}[p_i N\ell(z;x)] = \mathbb{E}_{z\sim\mathcal{P}}[\ell(z;x)].$$

Then we may redefine $f_i(x) = \mathbb{E}_{z\sim\mathcal{P}}[\ell(z;x)]$, while sampling from $\mathcal{P}$ in FedAvg can be approximated by sampling from $\mathcal{P}_i$. This setting is usually referred to as the *IID setting* in federated learning literature, and the resulting FedAvg is sometimes also called *local SGD* or *parallel SGD*. Note that for FedAvg in the IID setting, as long as the total number of participating clients S is fixed, the detailed client selection strategy does not make a difference.

The convergence of FedAvg in the IID setting (or local/parallel SGD) has been investigated in a series of papers including [Zhou and Cong, 2018], [Stich, 2019], [Woodworth et al., 2020], [Wang and Joshi, 2021], etc. The following result was established in [Woodworth et al., 2020].

Theorem 7.1. *Suppose $f(x) = \mathbb{E}_{z\sim\mathcal{P}}[\ell(z;x)]$ is convex and L-smooth, and has a minimizer $x^* \in \mathbb{R}^d$. Assume that*

$$\mathbb{E}_{z\sim\mathcal{P}}\left[\|\nabla_x\ell(z;x) - \nabla f(x)\|^2\right] \le \sigma^2$$

for some $\sigma > 0$, and that $\|x(0) - x^*\| \le R$ for some $R > 0$. Then for $\alpha = 1$ and $\eta \le 1/(4L)$, we have

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \le \frac{2R^2}{\eta KT} + \frac{2\eta\sigma^2}{BS} + \frac{8L(S-1)(K-1)\eta^2\sigma^2}{BS},$$

where

$$\bar{x}_T \coloneqq \frac{1}{SKT} \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} \sum_{i \in \mathcal{S}(t)} x_i^t(k).$$

Furthermore, by choosing the step size η properly, we can achieve

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq O\left(\frac{R\sigma/\sqrt{B}}{\sqrt{SKT}} + \frac{(L\sigma^2 R^4/B)^{1/3}}{K^{1/3}T^{2/3}} + \frac{LR^2}{KT} \right).$$

Since each outer iteration of FedAvg consists of one round communication between the server and the clients, we may count the number of outer iterations T needed to obtain an approximate optimal solution to characterize the communication complexity of FedAvg. By Theorem 7.1, it can be shown that, for an arbitrary $\epsilon > 0$, in order to achieve $\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \epsilon$, the communication complexity can be upper bounded by

$$O\left(\frac{\sigma^2/B}{SK\epsilon^2} + \frac{\sqrt{L}\sigma/\sqrt{B}}{K^{1/2}\epsilon^{3/2}} + \frac{L}{K\epsilon} \right).$$

It can be seen that, allowing more local SGD updates K for each participating client will indeed help reduce the communication complexity of FedAvg.

Connection with peer-to-peer consensus optimization. Note that the iterations of FedAvg in the IID setting with $\alpha = 1$ can be equivalently expressed as

$$x_i(t+1) = \begin{cases} \frac{1}{S} \sum_{j=1}^{S} \left(x_j(t) - \eta g_j(t) \right), & K \text{ divides } t+1, \\ x_i(t) - \eta g_i(t), & \text{otherwise}, \end{cases}$$

where $g_i(t)$ is a stochastic gradient of f_i at $x_i(t)$. By introducing the weight matrices

$$W(t) = \begin{cases} \frac{1}{S}\mathbf{1}\mathbf{1}^\mathsf{T}, & K \text{ divides } t+1, \\ I, & \text{otherwise}, \end{cases}$$

we can get

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

which is equivalent to the diffusion version of the decentralized stochastic gradient descent (DSGD) iterations with time-varying weight matrices. In addition, it's not hard to see that the sequence of weight matrices $(W(t))_{t \in \mathbb{N}}$ is K-strongly connected (see Chapter 6). This connection in the underlying mathematical structure suggests us that we may analyze FedAvg by using tools from DSGD with time-varying weight matrices. In fact, similar ideas were employed in [Wang and Joshi, 2021] that proposed a unified framework for local-update SGD algorithms.

Distinct Local Distributions

The convergence analysis of FedAvg when local distributions $\mathcal{P}_i$ are distinct is even more challenging, and currently there are still related questions that do not yet have satisfactory answers. The following result is from [Karimireddy et al., 2020].

Theorem 7.2. *Suppose each f_i is L-smooth and convex, and $f(x)$ has a minimizer x^*. Assume that there exists $\sigma > 0$ such that*

$$\mathbb{E}_{z\sim\mathcal{P}_i}\left[\|p_i N\nabla_x \ell(z;x) - \nabla f_i(x)\|^2\right] \le \sigma^2,$$

and that $\|x(0) - x^*\| \le R$ for some $R > 0$. Furthermore, suppose there exist $\delta > 0$ and $\beta > 0$ such that

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x)\|^2 \le \delta^2 + \beta^2\|\nabla f(x)\|^2. \tag{7.4}$$

Then by letting $\alpha \ge \sqrt{S}$ and $\eta \le 1/(8\alpha LK(1+\beta^2))$, we have

$$\mathbb{E}[f(\overline{x}_T) - f(x^*)] \le \frac{3R^2}{\alpha\eta TK} + \frac{\alpha\eta}{S}\left[\frac{\sigma^2}{B} + 2K\left(1 - \frac{S}{N}\right)\delta^2\right] + 54LK^2\alpha^2\eta^2\delta^2,$$

where $\overline{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x(t)$. Furthermore, by choosing the step sizes α and η properly, we can achieve

$$\mathbb{E}[f(\overline{x}_T) - f(x^*)] \le O\left(\frac{R}{\sqrt{SKT}}\sqrt{\frac{\sigma^2}{B} + 2K\left(1 - \frac{S}{N}\right)\delta^2} + \frac{(L\delta^2R^4)^{1/3}}{(T+1)^{2/3}} + \frac{\beta^2LR^2}{T}\right).$$

The condition (7.4) in Theorem 7.2 is called (δ,β)-*bounded gradient dissimilarity* by [Karimireddy et al., 2020]. Some special cases when this condition holds include:

- In the IID case where $f_i(x) = f(x)$ for all i, we have $\delta = 0$ and $\beta = 1$.
- When the local gradients $\nabla f_i(x)$ satisfy

  $$\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x) - \nabla f(x)\|^2 \le \tilde{\delta}^2$$

  for some $\tilde{\delta} > 0$, we then have

  $$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x)\|^2 &\le \frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x) - \nabla f(x) + \nabla f(x)\|^2\\ &\le \frac{2}{N}\sum_{i=1}^{N}\left(\|\nabla f_i(x) - \nabla f(x)\|^2 + \|\nabla f(x)\|^2\right) \le 2\tilde{\delta}^2 + 2,\end{aligned}$$

  which gives $\delta = \sqrt{2}\tilde{\delta}$ and $\beta = \sqrt{2}$.
- When each f_i is L-smooth and μ-strongly convex, denote

  $$\Delta = f(x^*) - \frac{1}{N}\sum_{i=1}^{N}\min_x f_i(x).$$

Then we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x)\|^2 &\le \frac{2L}{N}\sum_{i=1}^{N}\left(f_i(x)-\min_y f_i(y)\right) = 2L\left(f(x)-\frac{1}{N}\sum_{i=1}^{N}\min_y f_i(y)\right)\\ &= 2L(f(x)-f(x^*))+2L\Delta\\ &\le \frac{L}{\mu}\|\nabla f(x)\|^2+2L\Delta,\end{aligned}$$

where the first step follows from $\|\nabla h(x)\|^2 \le 2L(h(x)-\inf_y h(y))$ for any lower-bounded L-smooth function h, and the last step follows from $h(y) \le h(x)+\langle\nabla h(x), y-x\rangle+\frac{1}{2\mu}\|\nabla h(y)-\nabla h(x)\|^2$ for continuously differentiable and μ-strongly convex h (see [Nesterov, 2018, Theorem 2.1.10]). We get $\delta=\sqrt{2L\Delta}$ and $\beta=\sqrt{L/\mu}$.

The above cases indicate that the pair (δ,β) indeed quantifies the dissimilarities between the local cost functions f_i.

By Theorem 7.2, we can derive the communication complexity bound of FedAvg with distinct $\mathcal{P}_i$ given by

$$O\left(\frac{\sigma^2/B}{SK\epsilon^2}+\left(1-\frac{S}{N}\right)\frac{\delta^2}{S\epsilon^2}+\frac{\sqrt{L}\delta}{\epsilon^{3/2}}+\frac{\beta^2 L}{\epsilon}\right).$$

Here we keep the constants δ and β in the complexity bound to illustrate how the dissimilarities between f_i affect the complexity; also note that the quantities L and σ may become worse if some p_i is much larger than $1/N$. By comparing the communication complexity bound with the IID case, we can see that

1. As the local distributions $\mathcal{P}_i$ and the proportions p_i differ more among the clients, the communication complexity will in general become worse. While this is not a surprising result, it demonstrates that local datasets being non-IID can indeed lead to challenges in federated learning.
2. There is only one term in the communication complexity bound that will vanish as $K\to\infty$, and the other three terms are independent of K. As a result, the benefit of increasing the number of local updates K will ultimately vanish. This is different from the IID setting where by increasing K we can decrease the communication complexity bound to zero.

Notes on References

This chapter only introduces federated learning from a distributed optimization viewpoint, and many other interesting and important aspects of federated learning have not been covered. For a relatively recent introduction and survey to the broad topic of federated learning, we recommend the paper [Kairouz et al., 2021].

Bibliography

[Kairouz et al., 2021] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

[Karimireddy et al., 2020] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143.

[McMahan et al., 2017] McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282.

[Nesterov, 2018] Nesterov, Y. (2018). *Lectures on Convex Optimization*. Springer, 2 edition.

[Stich, 2019] Stich, S. U. (2019). Local SGD converges fast and communicates little. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

[Wang and Joshi, 2021] Wang, J. and Joshi, G. (2021). Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms. *The Journal of Machine Learning Research*, 22(1):9709–9758.

[Woodworth et al., 2020] Woodworth, B., Patel, K. K., Stich, S., Dai, Z., Bullins, B., Mcmahan, B., Shamir, O., and Srebro, N. (2020). Is local SGD better than minibatch SGD? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10334–10343.

[Zhou and Cong, 2018] Zhou, F. and Cong, G. (2018). On the convergence properties of a K-step averaging stochastic gradient descent algorithm for nonconvex optimization. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3219–3227.