

Stochastic Optimization (Module 1: Intro)

Why optimization?

Optimization is an multidisciplinary field and it is ubiquitous in industry and real-life decision making. Optimization problems arise in many areas of engineering and science. For example:

- **Industrial Engineering:** logistics, scheduling, manufacturing, queuing, facility layout, healthcare
- **Chemical Engineering:** in minimizing energy consumption, optimization of the separation process
- **Computer Engineering:** artificial intelligence, specifically machine learning and big data applications
- **Electrical Engineering:** image/signal processing, wireless communication networks, sensor networks, and social networks; power systems and market
- **Mechanical Engineering:** in design, e.g., engines, optimal weight design of a gear train, process parameter optimization in casting

What is an optimization problem?

In optimization, given a set $X \subseteq \mathbb{R}^n$ and a function $f : X \rightarrow \mathbb{R}$, the goal is to solve the following problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X. \end{array} \quad (P)$$

By solving problem (P), we mean to seek a vector $x^* \in X$ such that $f(x^*) \leq f(x)$ for all $x \in X$.

Definition (optimal solution): A vector $x^* \in X$ is called an optimal solution to problem (P) if $f(x^*) \leq f(x)$ for all $x \in X$.

Definition (optimal objective value): The lowest value of $f(x)$ over all $x \in X$ is called the optimal objective value.

What is a stochastic optimization (SO) problem?

Example (a motivating application of SO in IE): A firm has N distribution centers (DCs) and M retail locations (RLs). For the next year, the firm needs to decide how much of the product to stock at the DCs and how much to ship to the RLs.

Indices:

- $i \in \{1, \dots, N\}$ denotes the i th DC
- $j \in \{1, \dots, M\}$ denotes the j th RL

Problem parameters (could be known or afflicted with uncertainty):

- s_i is per unit stocking cost at DC i
- c_{ij} is per unit shipping cost from i th DC to j th RL
- d_j is demand at j th retail location.

Decision variables:

- z_i denotes the amount of products stored at the i th DC
- y_{ij} denotes the amount of products shipped from the i th DC to the j th RL

Deterministic model:

If the parameters including the demand is known, then we can consider the following optimization problem:

$$\begin{array}{ll} \text{minimize}_{z,y} & \sum_{i=1}^N s_i z_i + \sum_{i=1}^N \sum_{j=1}^M c_{ij} y_{ij} \\ \text{subject to} & \\ & \sum_{j=1}^M y_{ij} \leq z_i \quad \text{for all } i \in \{1, \dots, N\} \\ & \sum_{i=1}^N y_{ij} \geq d_j \quad \text{for all } j \in \{1, \dots, M\} \\ & z_i \geq 0 \quad \text{for all } i \in \{1, \dots, N\} \\ & y_{ij} \geq 0 \quad \text{for all } i \in \{1, \dots, N\} \text{ and all } j \in \{1, \dots, M\}. \end{array}$$

Stochastic formulations

Now suppose that the demand at each retail location is uncertain.

Generally, we can view each d_j as a **random variable** that may take values within a range. Let \tilde{d}_j denote the random variable for the demand at the j th RL.

About \tilde{d}_j : It has an unknown probability distribution. We can use historical data to fit a probability distribution. For example, \tilde{d}_j may have a normal distribution with mean μ and variance σ^2 .

Approach 1 (Expected-valued formulation):

Suppose it is possible to violate the demand constraints, but there is a stockout penalty defined as follows:

- p_j is the stockout penalty at the j th RL

The firm wants to make its stocking and shipping decisions in a way to minimize the total costs. We can consider the following formulation.

$$\begin{aligned} &\text{minimize}_{z,y} \quad \sum_{i=1}^N s_i z_i + \sum_{i=1}^N \sum_{j=1}^M c_{ij} y_{ij} + \mathbb{E} \left[\sum_{j=1}^M p_j \max \{0, \tilde{d}_j - \sum_{i=1}^N y_{ij}\} \right] \\ &\text{subject to} \\ &\quad \sum_{j=1}^M y_{ij} \leq z_i \quad \text{for all } i \in \{1, \dots, N\} \\ &\quad z_i \geq 0 \quad \text{for all } i \in \{1, \dots, N\} \\ &\quad y_{ij} \geq 0 \quad \text{for all } i \in \{1, \dots, N\} \text{ and all } j \in \{1, \dots, M\}. \end{aligned}$$

Approach 2 (Chance constrained formulation):

Suppose the firm wants to make its stocking and shipping decisions in a way to guarantee with a probability of $1 - \epsilon$ that the demand will be satisfied. We can consider the following formulation.

$$\begin{aligned} &\text{minimize}_{z,y} \quad \sum_{i=1}^N s_i z_i + \sum_{i=1}^N \sum_{j=1}^M c_{ij} y_{ij} \\ &\text{subject to} \\ &\quad \sum_{j=1}^M y_{ij} \leq z_i \quad \text{for all } i \in \{1, \dots, N\} \\ &\quad \text{Prob} \left(\tilde{d}_j - \sum_{i=1}^N y_{ij} \leq 0 \text{ for all } j \in \{1, \dots, M\} \right) \geq 1 - \epsilon \\ &\quad z_i \geq 0 \quad \text{for all } i \in \{1, \dots, N\} \\ &\quad y_{ij} \geq 0 \quad \text{for all } i \in \{1, \dots, N\} \text{ and all } j \in \{1, \dots, M\}. \end{aligned}$$

Reading assignment:

If you need to refresh your memory about random variables and expectations, read Chapters 4 and 5 of the following book.

- Applied Statistics and Probability for Engineers, 6th Edition, by Douglas C. Montgomery and George C. Runger

More formal formulations for stochastic optimization

Expected-valued model: The aforementioned model is an example of the following formulation

$$\begin{aligned} &\text{minimize}_x \quad f(x) \triangleq \mathbb{E}[F(x, \xi)] \\ &\text{subject to} \\ &\quad x \in X. \end{aligned}$$

(SO)

- Here $\mathbb{E}[\bullet]$ denotes the expectation of a random variable with respect to the random variables.
- $x \in \mathbb{R}^n$ is the vector of decision variables,
- $\xi \in \mathbb{R}^d$ denotes the random variables,
- $F : \mathbb{R}^n \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a stochastic function,
- and $X \subseteq \mathbb{R}^n$ is the constraint set.

Chance constrained model: The aforementioned model is an example of the following formulation

$$\begin{array}{ll} \text{minimize}_x & F(x) \\ \text{subject to} & \mathbb{P}[G(x, \xi) \leq 0] \geq 1 - \epsilon \\ & x \in X. \end{array}$$

(CC)

In-class assignment 1: Consider model (SO). Determine the following terms in the aforementioned example.

- $x = (z, y) = (z_1, \dots, z_N, y_{11}, \dots, y_{1M}, y_{21}, \dots, y_{2M}, \dots, y_{N1}, \dots, y_{NM})$
- $\xi = (\tilde{d}_1, \dots, \tilde{d}_M)$
- $F(x, \xi) = \sum_{i=1}^N s_i z_i + \sum_{i=1}^N \sum_{j=1}^M c_{ij} y_{ij} + \sum_{j=1}^M p_j \max \{0, \tilde{d}_j - \sum_{i=1}^N y_{ij}\}$
- $X = \{(z, y) \mid \sum_{j=1}^M y_{ij} \leq z_i \text{ for all } i \in \{1, \dots, N\}; z_i, y_{ij} \geq 0 \text{ for all } i \in \{1, \dots, N\} \text{ and all } j \in \{1, \dots, M\}\}$

Why are the above formulations challenging to solve?

What is the expectation of a function of a random variable? What about probability?

Recall that given a random variable u and its density function $p(u)$, the expected value of u is defined as follows:

$$E(u) = \int_{-\infty}^{+\infty} up(u)du.$$

Now, consider $v = h(u)$ as a new random variable. Then the expected value of v is as follows

$$E(v) = \int_{-\infty}^{+\infty} h(u)p(u)du.$$

For this reason, we can write for every given $x \in X$,

$$\mathbb{E}[F(x, \xi)] = \int_{\mathbb{R}^d} F(x, \xi)p(\xi)d\xi.$$

Also, we can write for every given $x \in X$,

$$\mathbb{P}[G(x, \xi) \leq 0] = \int_{\{\xi \mid G(x, \xi) \leq 0\}} p(\xi)d\xi.$$

The two formulations are challenging because

- Often, the density function $p(\xi)$ is unknown and we only have access to the past data, i.e., samples of the random variable observed in the past.
- Generally, evaluation of a multi-dimensional integral takes much time, in particular, when the dimension of the random variable is high.
- In addition, when the model becomes nonlinear, it is generally more difficult to solve optimization problems.

Some basic notation and definitions

Vectors

A real n -dimensional vector is an ordered set of n real numbers $\{x_1, x_2, \dots, x_n\}$ and is usually written in the form of a column vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

The numbers x_1, x_2, \dots, x_n are called the *components* of x .

- The set of all n -dimensional vectors is called **Euclidean space** and is usually denoted by \mathbb{R}^n .

Matrices

A real matrix is a rectangular array of real numbers composed of rows and columns. We write

$$A = [a_{ij}]_{m \times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

- For a matrix of m rows and n columns, and we say that the matrix A is of *order* $m \times n$.
- We will deal with real matrices (i.e. $A \in \mathbb{R}^{m \times n}$) and $m \times n$ will always denote the rows \times columns.
- Given a matrix $A \in \mathbb{R}^{m \times n}$, the **transpose** of A is an $n \times m$ matrix whose rows are the columns of A . The transpose matrix of A is denoted by A^T .

Linear functions

A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ in the form

$$f(x) = c^T x = c_1 x_1 + c_2 x_2 + \dots + c_n x_n,$$

where

- $c_j, j = 1, \dots, n$ are constants and
- $x_j, j = 1, \dots, n$ are variables,

is called a *linear function*.

The notion of nonlinearity

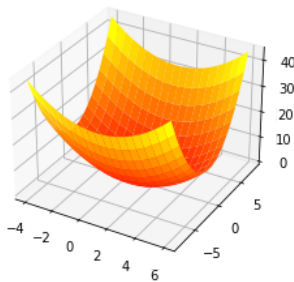
Examples of quadratic functions: Let function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined as

$$f(x_1, x_2) \triangleq \frac{1}{2} (a x_1^2 + b x_2^2) - x_1,$$

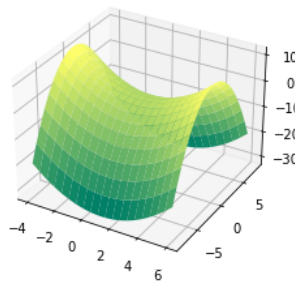
where $a, b \in \mathbb{R}$ are given parameters. Let us consider two instances:

```
In [27]: 1 import numpy as np
2 import matplotlib.pyplot as plt
3 from mpl_toolkits import mplot3d
4
5 def q_function(a,b,x1, x2):
6     return (0.5*a*x1**2 + 0.5*b*x2**2-x1)
7
8 def plt_title(a,b):
9     return '$f(x_1,x_2)=0.5(a\ x_1^2+ b\ x_2^2)-x_1$' for a= '+str(a)+' , b= '+str(b)
10
11 fig, [ax1, ax2] = plt.subplots(1, 2, figsize=(12,4),
12                               subplot_kw={'projection': '3d'})
13
14 x = np.linspace(-4, 6, 20)
15 y = np.linspace(-8, 8, 20)
16 X, Y = np.meshgrid(x, y)
17
18 a,b = 1,1
19 Z = q_function(a,b,X, Y)
20 ax1.plot_surface(X, Y, Z, cmap='autumn')
21 ax1.set_title(plt_title(a,b))
22
23
24 a,b = 1,-1
25 Z = q_function(a,b,X, Y)
26 ax2.plot_surface(X, Y, Z, cmap='summer')
27 ax2.set_title(plt_title(a,b))
28
29
30 #plt.savefig("QP_examples.pdf", dpi=300)
31 plt.show()
```

$$f(x_1, x_2) = 0.5(a x_1^2 + b x_2^2) - x_1 \quad \text{for } a = 1, b = 1$$



$$f(x_1, x_2) = 0.5(a x_1^2 + b x_2^2) - x_1 \quad \text{for } a = 1, b = -1$$

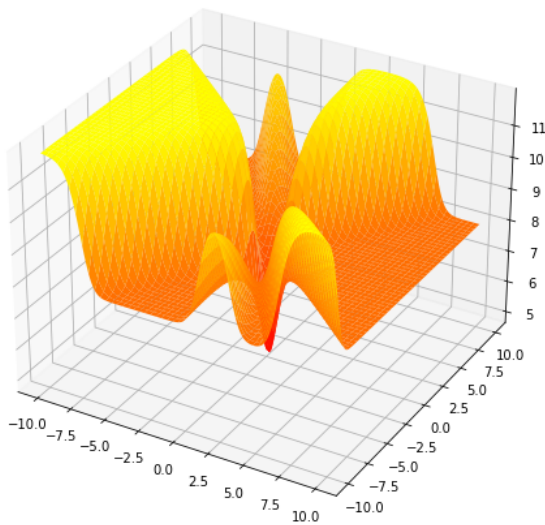


A nonlinear function that arises from machine learning:

```

In [26]: 1 import numpy as np
2 import matplotlib.pyplot as plt
3 from mpl_toolkits import mplot3d
4 from numpy import linalg as LA
5 import time
6 from math import *
7 import warnings
8 warnings.filterwarnings("ignore")
9
10 y = [1.165,0.626,0.075,0.351,-0.696]
11 z = [1,-1,-1,1,1]
12
13 def obj_f(u0,u1,y,z):
14     u = np.array([u0,u1])
15     output_list = [(z[i]-h_f(u,y[i]))**2 for i in range(len(y))]
16     return sum(output_list)
17
18 def phi_f(t):
19     return (np.exp(t)-np.exp(-t))/(np.exp(t)+np.exp(-t))
20
21 def h_f(u,y):
22     return phi_f(u[0]+u[1]*y)
23
24 fig = plt.figure(figsize=(12,8))
25 ax = fig.gca(projection='3d')
26
27 u0 = np.linspace(-10, 10, 500)
28 u1 = np.linspace(-10, 10, 500)
29 U0, U1 = np.meshgrid(u0, u1)
30
31
32 Z = obj_f(U0, U1,y,z)
33 ax.plot_surface(U0, U1, Z, cmap='autumn')
34
35 plt.show()

```



Stochastic optimization problems in machine learning

- Classification, clustering, regression, ...

$$\text{minimize}_x \quad \mathbb{E}_{(u,v)} [\mathcal{L}(h(u, x), v)]$$

$$\text{subject to} \quad x \in \mathbb{R}^n.$$

Here $h : \mathbb{R}^{p \times n} \rightarrow \mathbb{R}$ and $v \in \mathbb{R}$ and $\mathcal{L} : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a loss function.

Examples of the loss function \mathcal{L} :

- $\mathcal{L}(z, b) := \log(1 + \exp(-bz))$ in logistic regression
- $\mathcal{L}(z, b) := \max\{0, 1 - bz\}$ in linear support vector machines (SVM)

Empirical risk minimization (sample average approximation of the above problem)

$$\text{minimize}_x \quad \frac{1}{|S|} \sum_{\ell \in S} \mathcal{L}(h(u_\ell, x), v_\ell)$$

$$\text{subject to} \quad x \in \mathbb{R}^n.$$

- The dataset $D \triangleq \{(u_\ell, v_\ell) \mid \ell \in S\}$
- $|S|$ denotes the number of elements in the set S

Norm

Norm of a vector $x \in \mathbb{R}^n$ is a real-valued function and is denoted by $\|x\|$. It has the following properties:

- $\|x\| \geq 0$ for all $x \in \mathbb{R}^n$.
- $\|\lambda x\| = |\lambda| \cdot \|x\|$ for any $\lambda \in \mathbb{R}$ and any $x \in \mathbb{R}^n$.
- $\|x\| = 0$ iff $x = 0$.
- It satisfies the **triangle inequality**, i.e., $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$.

An example of norm, is the family of ℓ_p -norm defined as

$$\|x\|_p \triangleq \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}},$$

where $p \geq 1$. Particularly,

when $p = 1$, ℓ_1 -norm is given by $\|x\|_1 \triangleq \sum_{i=1}^n |x_i|$,

when $p = 2$, ℓ_2 -norm **Euclidean norm** is given by $\|x\|_2 \triangleq \sqrt{\sum_{i=1}^n |x_i|^2} = \sqrt{x^T x}$.

The ℓ_∞ -norm (maximum norm) is defined by $\|x\|_\infty \triangleq \max_{i=1, \dots, n} |x_i|$.

Throughout the course, unless otherwise specified, we use $\|\cdot\|$ to denote the Euclidean norm.

In-class assignment 2: Consider the Euclidean norm. Use the definition of norm provided earlier to show that the Euclidean norm is a norm.

Gradient mapping

Let function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be given. Assuming that all the partial derivatives of f exist, the gradient of f at $x \in \mathbb{R}^n$ is defined as

$$\nabla f(x) \triangleq \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

Example (gradient of a linear function): Let $f(x) = c^T x + d$ for any $x \in \mathbb{R}^n$, where $c \in \mathbb{R}^n$ and $d \in \mathbb{R}$ are known parameters. Find $\nabla f(x)$.

Solution:

$$f(x) = c^T x + d = [c_1, \dots, c_n] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + d = d + \sum_{i=1}^n c_i x_i.$$

We then can write:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial (d + \sum_{i=1}^n c_i x_i)}{\partial x_1} \\ \vdots \\ \frac{\partial (d + \sum_{i=1}^n c_i x_i)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = c.$$

Example (gradient of norm squared): Let $f(x) = \|x\|_2^2$ for any $x \in \mathbb{R}^n$.

Solution:

$$f(x) = x^T x = [x_1, \dots, x_n] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^n x_i^2.$$

We then can write:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial \sum_{i=1}^n x_i^2}{\partial x_1} \\ \vdots \\ \frac{\partial \sum_{i=1}^n x_i^2}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ \vdots \\ 2x_n \end{bmatrix} = 2x.$$

Hessian matrix

Let function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be given. Assuming that all the second-order partial derivatives of f exist, the Hessian of f at $x \in \mathbb{R}^n$ is defined as

$$\nabla^2 f(x) \triangleq \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right]_{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}.$$

In-class assignment 3:

- Suppose $f(x) = \frac{1}{2} \|x\|_2^2$. Show that $\nabla^2 f(x) = \mathbf{I}_n$ where \mathbf{I}_n denotes the identity matrix of size n .

Out-class assignment:

- Let $f(x) = \frac{1}{2} x^T Q x$ for any $x \in \mathbb{R}^n$, where $Q \in \mathbb{R}^{n \times n}$ is a given square symmetric matrix. Show that we have for any $x \in \mathbb{R}^n$

$$\nabla f(x) = Qx \quad \text{and} \quad \nabla^2 f(x) = Q.$$

First approach:

$$x^T Q x = [x_1, \dots, x_n] \begin{bmatrix} Q_{1,1} & \dots & Q_{1,n} \\ \vdots & \ddots & \vdots \\ Q_{n,1} & \dots & Q_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^n \sum_{j=1}^n Q_{i,j} x_i x_j = \dots$$

Second approach:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_i e_i + \begin{bmatrix} x_1 \\ \vdots \\ 0 \\ \vdots \\ x_n \end{bmatrix} = x_i e_i + x_{-i}$$

$$x^T Q x = (x_i e_i + x_{-i})^T Q (x_i e_i + x_{-i}) = (x_i e_i^T Q + x_{-i}^T Q) (x_i e_i + x_{-i})$$

$$= x_i e_i^T Q x_i e_i + x_i e_i^T Q x_{-i} + x_i x_{-i}^T Q e_i + x_{-i}^T Q x_{-i}$$

$$= (e_i^T Q x_i e_i) x_i^2 + 2 (e_i^T Q x_{-i}) x_i + x_{-i}^T Q x_{-i}.$$

$$\frac{\partial (x^T Q x)}{\partial x_i} = 2 e_i^T Q e_i x_i + 2 e_i^T Q x_{-i}$$

$$= 2 e_i^T Q (e_i x_i + x_{-i})$$

$$= 2 e_i^T Q x$$

$$= 2 Q_{\bullet, i} x.$$

Local optimal vs global optimal solution

Let the decision vector x denote an n -tuple of real numbers (x_1, \dots, x_n) . Consider a constraint set $X \subseteq \mathbb{R}^n$, and a function $f(x)$ where $f : X \rightarrow \mathbb{R}$.

The goal in optimization is to find an $x^* \in X$ such that

$$f(x^*) \leq f(x), \quad \text{for all } x \in X.$$

Continuous optimization problems: analysis is done using the mathematics of calculus and convexity.

- Nonlinear programming: the case where either f is nonlinear or X is specified by nonlinear equations or inequalities. This is the focus of the course.

Discrete optimization problems: analysis is done using combinatorial and discrete mathematics.

This first topic is focused on **unconstrained differentiable nonlinear optimization** given as:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathbb{R}^n. \end{array}$$

A vector x^* is an **unconstrained global minimum** of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ if

$$f(x^*) \leq f(x), \quad \text{for all } x \in \mathbb{R}^n.$$

A vector x^* is an **unconstrained local minimum** of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ if there is an $\epsilon > 0$ such that

$$f(x^*) \leq f(x), \quad \text{for all } x \in \mathbb{R}^n \text{ with } \|x - x^*\| \leq \epsilon.$$

A vector x^* is an **unconstrained local strict minimum** of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ if there is an $\epsilon > 0$ such that

$$f(x^*) < f(x), \quad \text{for all } x \in \mathbb{R}^n \text{ with } \|x - x^*\| \leq \epsilon \text{ and } x \neq x^*.$$

A vector x^* is a **constrained global minimum** of a function $f : X \rightarrow \mathbb{R}$ if

$$f(x^*) \leq f(x), \quad \text{for all } x \in X.$$

A vector x^* is a **constrained local minimum** of a function $f : X \rightarrow \mathbb{R}$ if there is an $\epsilon > 0$ such that

$$f(x^*) \leq f(x), \quad \text{for all } x \in X \text{ with } \|x - x^*\| \leq \epsilon.$$

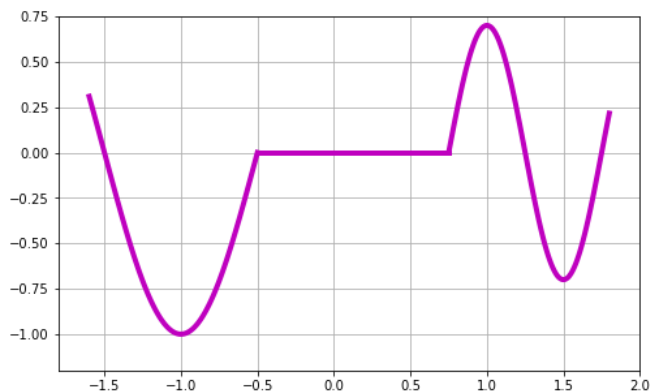
A vector x^* is a **constrained local strict minimum** of a function $f : X \rightarrow \mathbb{R}$ if there is an $\epsilon > 0$ such that

$$f(x^*) < f(x), \quad \text{for all } x \in X \text{ with } \|x - x^*\| \leq \epsilon \text{ and } x \neq x^*.$$


```

In [28]: 1 import matplotlib.pyplot as plt
          2 import numpy as np
          3
          4 #plt.style.use('fivethirtyeight')
          5 plt.figure(figsize=(8,5))
          6 plt.grid(True)
          7
          8 break_points = [-1.6,-0.5,0.75,1.8]
          9 cst = 1
         10 cst2 = 2
         11 line_width = 4
         12 plot_color = 'm'
         13 increm = 0.01
         14 theta = 0.7
         15
         16
         17 t1 = np.arange(break_points[0], break_points[1]+increm, increm)
         18 t1 = np.asarray(list(np.round(t1,2))[0:-1])
         19 s1 = np.cos(cst*np.pi*t1)
         20
         21 t2 = np.arange(break_points[1], break_points[2]+increm, increm)
         22 t2 = np.round(t2,2)
         23 s2 = np.array([np.cos(cst*np.pi*list(t2)[0]) for i in range(len(t2))])
         24
         25
         26 t3 = np.arange(break_points[2], break_points[3]+increm, increm)
         27 s3 = theta*(np.cos(cst2*np.pi*t3)-np.cos(cst2*np.pi*list(t3)[0])) + list(s2)[0]
         28
         29 plt.plot(t1, s1, lw=line_width, c= plot_color)
         30 plt.plot(t2, s2, lw=line_width, c= plot_color)
         31 plt.plot(t3, s3, lw=line_width, c= plot_color)
         32
         33 plt.xlim(break_points[0]-.2, break_points[3]+.2)
         34 plt.ylim(-1.2, 0.75)
         35
         36 plt.show()

```



Consider the above function over the set $X = [-1.6, 1.8]$. We have:

- Global minimum: $x = -1$
- Local minima: $x \in \{-1, 1.5\} \cup (-0.5, 0.75]$
- Strict local minima: $x \in \{-1, 1.5\}$

In-class assignment 4: Fill out the blanks below about the function in the plot above.

- Global maximum: $x = 1$
- Local maxima: $x \in \{-1.6, 1, 1.8\} \cup [-0.5, 0.75)$
- Strict local maxima: $x \in \{-1.6, 1, 1.8\}$
- Global maximum objective value around 0.70

Convex sets and convex functions

A subset $C \subseteq \mathbb{R}^n$ is called convex if for any $\alpha \in [0, 1]$ and any $x, y \in C$, we have

$$\alpha x + (1 - \alpha)y \in C.$$



One popular convex set in optimization is $C = \{x \in \mathbb{R}^n \mid Ax = b, x \geq 0\}$.

Function $f : C \rightarrow \mathbb{R}$ is called convex if:

- (a) The set C is convex;
- (b) For any $\alpha \in [0, 1]$ and any $x, y \in C$, we have

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

Function $f : C \rightarrow \mathbb{R}$ is called strictly convex if:

- (a) The set C is convex;
- (b) For any $\alpha \in (0, 1)$ and any $x, y \in C$ with $x \neq y$, we have

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

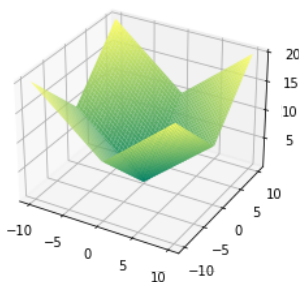
Examples of convex and concave functions:

- Exponential: e^{ax} is convex on \mathbb{R} for any $a \in \mathbb{R}$
- Powers: x^a is convex on \mathbb{R}_{++} when $a \geq 1$ or $a \leq 0$, and is concave if $0 \leq a \leq 1$
- Powers of absolute value: $|x|^p$ for $p \geq 1$, is convex on \mathbb{R}
- Logarithm: $\log x$ is concave on \mathbb{R}_{++}
- Negative entropy: $x \log x$ is convex on \mathbb{R}_{++}
- ℓ_p norm: $\|x\|_p$ for $p \geq 1$ is convex on \mathbb{R}^n
- Affine function: $f(X) = \text{tr}(A^T X) + b = \sum_{i=1}^m \sum_{j=1}^n A_{ij} X_{ij} + b$ where $A, X \in \mathbb{R}^{m \times n}$

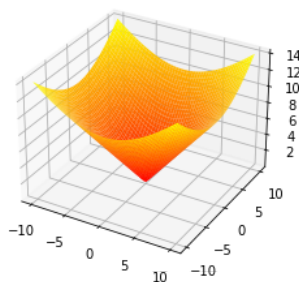
A poluar convex function is ℓ_p -norm for $p \geq 1$.

```
In [24]: 1 import numpy as np
2 import matplotlib.pyplot as plt
3 from mpl_toolkits import mplot3d
4
5 def norm_function(x1, x2, p):
6     return (abs(x1)**p + abs(x2)**p)**(1/p)
7
8 def norm_infinity(x1, x2):
9     return (abs(x1) + abs(x2) + abs(abs(x1) - abs(x2))) / 2
10
11 def plt_title(p):
12     return '$f(x) = \|x\|_p$ for p= ' + str(p)
13
14 fig, [ax1, ax2, ax3] = plt.subplots(1, 3, figsize=(16, 4),
15                                     subplot_kw={'projection': '3d'})
16
17 x = np.linspace(-10, 10, 100)
18 y = np.linspace(-10, 10, 100)
19 X, Y = np.meshgrid(x, y)
20
21 p=1
22 Z = norm_function(X, Y, p)
23 ax1.plot_surface(X, Y, Z, cmap='summer')
24 ax1.set_title(plt_title(p))
25 #ax1.view_init(-120, 60)
26
27 p=2
28 Z = norm_function(X, Y, p)
29 ax2.plot_surface(X, Y, Z, cmap='autumn')
30 ax2.set_title(plt_title(p))
31 #ax2.view_init(-120, 60)
32
33
34 Z = norm_infinity(X, Y)
35 ax3.plot_surface(X, Y, Z, cmap='winter')
36 ax3.set_title('$f(x) = \|x\|_p$ for p= $\\infty$')
37 #ax3.view_init(-120, 60)
38
39
40 plt.show()
```

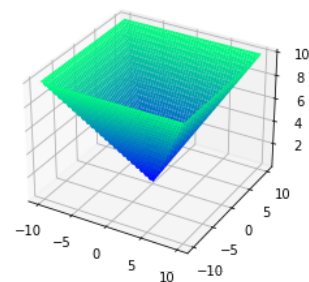
$f(x) = \|x\|_p$ for $p = 1$



$f(x) = \|x\|_p$ for $p = 2$



$f(x) = \|x\|_p$ for $p = \infty$



Examples of relations between norms in two dimensional space

Note that when $n = 2$, we have

$$\|x\|_1 = |x_1| + |x_2|$$

$$\|x\|_\infty = \max\{|x_1|, |x_2|\}$$

Thus, we have

$$\|x\|_\infty \leq \|x\|_1$$

$$\|x\|_1 \leq 2\|x\|_\infty$$

So,

$$0.5\|x\|_1 \leq \|x\|_\infty \leq \|x\|_1$$

Optimality conditions

Proposition 1: (Necessary optimality conditions for unconstrained nonconvex case)

Let x^* be an unconstrained local minimum of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ where f is continuously differentiable over an open set containing x^* . Then,

$$\nabla f(x^*) = 0_{n \times 1},$$

which is called *first-order necessary condition*.

If in addition, f is twice differentiable over the open set, then the Hessian matrix at x^* is positive semidefinite, i.e.,

$$\nabla^2 f(x^*) \succeq 0_{n \times n},$$

which is called *second-order necessary condition*.

Proposition 2: (Sufficient optimality conditions for unconstrained nonconvex case)

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable over an open set, which contains x^* such that

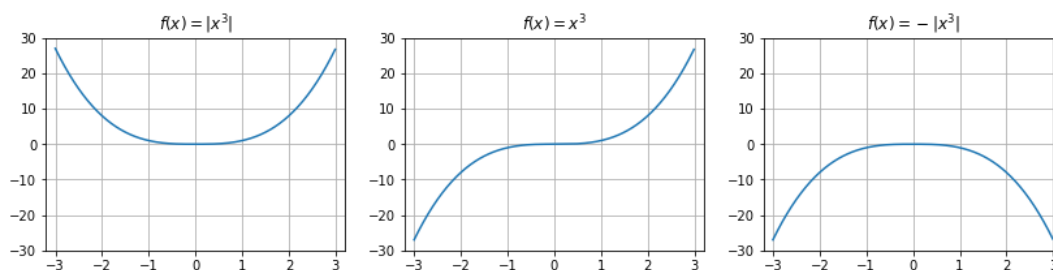
$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succ 0.$$

Then, x^* is a *strict unconstrained local minimum* of f . Moreover, there are scalars $\gamma > 0$ and $\epsilon > 0$ such that

$$f(x) \geq f(x^*) + \frac{\gamma}{2} \|x - x^*\|^2, \quad \text{for all } \|x - x^*\| < \epsilon.$$

Are the necessary conditions sufficient as well? The following counterexample shows they are not!

```
In [25]: 1 import matplotlib.pyplot as plt
2 import numpy as np
3 #plt.style.use('fivethirtyeight')
4
5 fig = plt.figure(figsize=(14,3))
6 increm=0.01
7 x = np.arange(-3, 3, increm)
8
9 plt.subplot(131)
10 y1 = np.absolute(x**3)
11 plt.plot(x, y1)
12 plt.xlim(-3.2, 3.2)
13 plt.ylim(-30, 30)
14 plt.title("$f(x) = |x^3|$")
15 plt.grid(True)
16
17 plt.subplot(132)
18 y2 = x**3
19 plt.plot(x, y2)
20 plt.xlim(-3.2, 3.2)
21 plt.ylim(-30, 30)
22 plt.title("$f(x) = x^3$")
23 plt.grid(True)
24
25 plt.subplot(133)
26 y3 = -np.absolute(x**3)
27 plt.plot(x, y3)
28 plt.xlim(-3.2, 3.2)
29 plt.ylim(-30, 30)
30 plt.title("$f(x) = -|x^3|$")
31 plt.grid(True)
32
33 plt.show()
```



In-class assignment 5:

Consider the problem of minimizing the function $f(x) = x^3$ over $x \in \mathbb{R}$.

- Is $x^* = 0$ a (local) optimal solution? No
- Calculate $\nabla f(x)$ and $\nabla^2 f(x)$. $\nabla f(x) = 3x^2$ and $\nabla^2 f(x) = 6x$
- Is the first-order necessary condition satisfied for $x^* = 0$? Yes
- Is the second-order necessary condition satisfied for $x^* = 0$? Yes
- Are the sufficient optimality conditions met for $x^* = 0$? No
- What do you conclude? That the necessary conditions are not sufficient.

Eigenvalues and eigenvectors

For an $n \times n$ matrix A , scalars λ and vectors $v \in \mathbb{R}^n$ satisfying:

$$Av = \lambda v$$

are called eigenvalues and eigenvectors of A , respectively, and any such pair, (λ, v) , is called an eigenpair of A .

About positive definiteness of a matrix

A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive semidefinite if all its eigenvalues are nonnegative, i.e.,

$$A \geq 0_{n \times n} \iff \min_{i \in \{1, \dots, n\}} \lambda_i(A) \geq 0.$$

A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive definite (denoted by $A > 0_{n \times n}$) if and only if

$$A \succ 0 \iff \min \lambda_i(A) > 0$$

How to evaluate eigenvectors and eigenvalues of a matrix in Python

```
In [ ]: 1 import numpy as np
2 from numpy import linalg as LA
3
4 my_matrix_1 = np.diag((2, -3, 6))
5 my_matrix_2 = np.array([[4, 1, -1], [1, 2, 1], [-1, 1, 2]])
6
7 print(my_matrix_1, "\n\n", my_matrix_2)
```

```
In [ ]: 1 w, v = LA.eig(my_matrix_1)
2 print(w)
```

```
In [ ]: 1 w, v = LA.eig(my_matrix_2)
2 print(w)
```

Review of linear algebra basics

Singular matrix: A matrix with a zero determinant is called singular. Otherwise, it is nonsingular or invertible.

The spectrum of A , denoted by $\sigma(A)$, is the set of distinct eigenvalues of A .

Proposition A.1: Let matrix A be $n \times n$. Then, the following hold:

- $\lambda \in \sigma(A) \iff A - \lambda I$ is singular $\iff \det(A - \lambda I) = 0$.
- $N(A - \lambda I)$ is called eigenspace of A . The set of $\{x \neq 0 \mid x \in N(A - \lambda I)\}$ is the set of all eigenvectors of A .
- Nonzero row vectors y^* such that $y^*(A - \lambda I) = 0$ are called left-hand eigenvectors of A .
- The eigenvalues of A are the solutions of the characteristic equation $p(\lambda) = 0$, where $p(\lambda) \triangleq \det(A - \lambda I)$ denotes the characteristic polynomial of A .
- Altogether, A has n eigenvalues, but some may be complex numbers and some may be repeated. Note that even if the entries of A are real numbers, some eigenvalues may be complex.
- $\text{trace}(A) \triangleq \sum_{i=1}^n a_{ii} = \sum_{i=1}^n \lambda_i$ where $\{\lambda_i\}_{i=1}^n$ denotes the set of eigenvalues of the square matrix A .
- $\det(A) = \prod_{i=1}^n \lambda_i$.
- The eigenvalues of a symmetric matrix A are all real numbers.
- The eigenvectors of a symmetric matrix A corresponding to different eigenvalues are orthogonal to each other.
- The eigenvalues of a triangular matrix are equal to its diagonal entries.
- The eigenvalues of A and A^T coincide.
- The eigenvalues of A^k are λ_i^k where $\{\lambda_i\}_{i=1}^n$ denotes the eigenvalues of A .
- If S is nonsingular, the eigenvalues of SAS^{-1} and A coincide.
- Eigenvalues of $A + cI$ are equal to $c + \lambda_i$ where $\{\lambda_i\}_{i=1}^n$ denotes the eigenvalues of A .
- A is singular if and only if it has an eigenvalue that is zero.
- Let A be symmetric. Let $\{\lambda_i\}_{i=1}^n$ and $\{v_i\}_{i=1}^n$ denote the eigenvalues and eigenvectors of A , respectively. Assume the eigenvalues are normalized, i.e., $\|v_i\|_2 = 1$ for $i = 1, \dots, n$. Then, we have

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

Positive definite matrices: A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is called positive definite and is denoted by $A \succ 0$ if $x^T A x > 0$ for all $x \in \mathbb{R}^n$, and $x \neq 0$.

- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive definite if and only if all eigenvalues of A are positive.

- If $A \in \mathbb{R}^{n \times n}$ is positive definite, and $H \in \mathbb{R}^{n \times n}$ is invertible, then $H A H^T$ is positive definite.

Positive semidefinite matrices: A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is called positive semidefinite and is denoted by $A \succeq 0$ if $x^T A x \geq 0$ for all $x \in \mathbb{R}^n$.

- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive semidefinite if and only if all eigenvalues of A are non-negative.
- If $A \in \mathbb{R}^{n \times n}$ is positive semidefinite, and $H \in \mathbb{R}^{m \times n}$ is a matrix, then $H A H^T$ is positive semidefinite.

Proposition A.2: Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Then,

$$\lambda_{\min} \|x\|_2^2 \leq x^T A x \leq \lambda_{\max} \|x\|_2^2, \quad \text{for all } x \in \mathbb{R}^n,$$

where λ_{\min} and λ_{\max} denote the minimum and maximum eigenvalues of A .

Proposition A.3: Let A be an $m \times n$ matrix. Then,

- $A^T A$ is symmetric and positive semidefinite.
- $A^T A$ is positive definite if and only if $\text{Rank}(A) = n$ and $m \geq n$.
- If $m = n$, $A^T A$ is positive definite if and only if A is nonsingular.

Matrix norm: Let A be an $n \times n$ matrix and $\|\cdot\|$ denote a vector norm. Then, matrix norm induced by the norm $\|\cdot\|$ is defined as

$$\|A\| \triangleq \sup_{\|x\|=1} \|Ax\|.$$

- Note that we have $\|A\| = \|A^T\|$.
- Let $A \in \mathbb{R}^{n \times n}$ and $x \in \mathbb{R}^n$. For any norm $\|\cdot\|$ and its induced matrix norm, we have

$$\|Ax\| \leq \|A\| \|x\|.$$

This inequality serves as an equivalence of Schwarz inequality for vectors.

Spectral radius of a matrix: Let A be an $n \times n$ matrix. The spectral radius of A is denoted by $\rho(A)$ and is defined as

$$\rho(A) \triangleq \max_{i=1, \dots, n} |\lambda_i|,$$

where $\{\lambda_i\}_{i=1}^n$ denotes the eigenvalues of A .

Proposition A.4:

Let A be an $n \times n$ matrix. For any induced matrix norm $\|\cdot\|$, we have

$$\rho(A) \leq \|A\|.$$

If A is symmetric, we have $\rho(A) = \|A\|_2 = \max\{|\lambda_{\min}|, |\lambda_{\max}|\}$, where $\|A\|_2$ denotes the matrix norm induced by Euclidean norm.

Stochastic Optimization (Module 2: Gradient Descent)

minimize $f(x)$ subject to: $x \in X$.

Notation: Throughout the course, often by writing $x \geq 0$ we mean x is an n -dimensional vector whose elements are all non-negative. Here, 0 is also an n -dimensional vector. We also use the notation $\mathbb{R}_{++} = \{x \in \mathbb{R}^n \mid x > 0\}$ and $\mathbb{R}_+ = \{x \in \mathbb{R}^n \mid x \geq 0\}$.

When does a feasible optimization problem have a solution?

Generally, an optimal solution may not exist. For example,

- minimization of $f(x) = x$ for $x \in \mathbb{R}$ does not have a solution
- minimization of $f(x) = e^x$ for $x \in \mathbb{R}$ does not have a solution
- minimization of $f(x) = -\log(x)$ for $x \in \mathbb{R}_{++}$ does not have a solution
- minimization of $f(x) = \frac{1}{x}$ for $x \in \mathbb{R}_{++}$. Here, even though $f^* = \inf_{x \in \mathbb{R}_{++}} \frac{1}{x} = 0$, does not have a solution. We say the optimal value is not **attained**.

Weierstrass' Theorem: Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a continuous function and $X \subset \mathbb{R}^n$ be a nonempty compact set. Then, both $\inf_{x \in X} f(x)$ and $\sup_{x \in X} f(x)$ are finite and are attained, i.e., minimizer in $\inf_{x \in X} f(x)$ and maximizer in $\sup_{x \in X} f(x)$ exist.

Examples of a nonconvex function

- Composition of two convex functions could be nonconvex. For example, suppose $g(x)$ is convex. The function $f(x) \triangleq (g(x))^2$ is not necessarily convex. For example, $f(x) = (x^2 - x)^2$ for $x \in \mathbb{R}$ is nonconvex.
- $f(x) = \frac{1}{2}x^T Qx$ when $Q \in \mathbb{R}^{n \times n}$ has some negative eigenvalues (see the examples in the last lecture).
- $f(x) = \|x\|_0$ over \mathbb{R}^n where $\|x\|_0$ denotes the number of nonzero elements in the vector x .

Proposition (Composition of convex functions): Let $h : \mathbb{R} \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}$ be given. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be defined as:

$$f(x) \triangleq h(g(x)).$$

Then, f is convex in the following two cases:

- g is convex, h is nondecreasing and convex
- g is concave, h is nonincreasing and convex

Definition (bounded set, closed set, compact set):

- A set $X \subseteq \mathbb{R}^n$ is called **bounded** if we have

$$\|x\| \leq M \quad \text{for all } x \in X \text{ for some } M \geq 0.$$
- A set $X \subseteq \mathbb{R}^n$ is called **closed** if its complement is open. Recall that $Y \subseteq \mathbb{R}^n$ is open if every point in Y is the center of an open ball contained in Y .
- A set $X \subseteq \mathbb{R}^n$ is **compact** if and only if it is closed and bounded.

Bounded mapping: Let $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a mapping. F is called bounded over $X \subseteq \mathbb{R}^n$ if we have

$$\|F(x)\| \leq M \quad \text{for all } x \in X \text{ for some } M \geq 0.$$

Definition (Strong convexity): Consider a continuously differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. We say f is μ -strongly convex over $X \subseteq \mathbb{R}^n$ if for some $\mu > 0$ we have

$$f(y) + \nabla f(y)^T (x - y) + \frac{\mu}{2} \|x - y\|^2 \leq f(x), \quad \text{for all } x, y \in X.$$

In-class assignment 1: Show that for a μ -strongly convex function f , we have

$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq \mu \|x - y\|^2, \quad \text{for all } x, y \in X.$$

$$f(y) + \nabla f(y)^T(x - y) + \frac{\mu}{2} \|x - y\|^2 \leq f(x).$$

$$f(x) - \nabla f(x)^T(x - y) + \frac{\mu}{2} \|x - y\|^2 \leq f(y).$$

$$\nabla f(y)^T(x - y) - \nabla f(x)^T(x - y) + \mu \|x - y\|^2 \leq 0$$

$$\mu \|x - y\|^2 \leq (\nabla f(x) - \nabla f(y))^T(x - y).$$

Remark: A strongly convex function is also a strictly convex function; and strictly convex function is also a convex function.

Proposition (Second derivative characterizations): Let $X \subseteq \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable over \mathbb{R}^n .

- (a) If $\nabla^2 f(x) \geq 0$ for all $x \in X$, then f is convex over X .
- (b) If $\nabla^2 f(x) > 0$ for all $x \in X$, then f is strictly convex over X .
- (c) If X is open and f is convex over X , then $\nabla^2 f(x) \geq 0$ for all $x \in X$.

The notion of L-smoothness

Definition (Lipschitz continuity): A differentiable function $f : X \rightarrow \mathbb{R}$ is called L -smooth if it has **Lipschitz continuous gradients** with parameter L , i.e., if there exists a scalar $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \quad \text{for all } x, y \in X.$$

- L -smoothness of a function, that is Lipschitz continuity of its gradients, is a stronger assumption than f being continuously differentiable.

Examples:

- $f(x) = x^2$ is 2-smooth over \mathbb{R} (why?). Is it 3-smooth as well? What do you conclude?

$$\|\nabla f(x) - \nabla f(y)\| = |2x - 2y| = 2|x - y| \leq 3|x - y|.$$

In-class assignment 2: Consider $f(x) = |x|$ for all $x \in \mathbb{R}$. Prove that f is not L -smooth.

$$|1 - (-1)| \leq L \|\epsilon - (-\epsilon)\| = 2L\epsilon \Rightarrow \epsilon \geq \frac{1}{L}.$$

Descent Lemma: Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be L -smooth over the set $X \subseteq \mathbb{R}^n$. Then, for any $x, y \in X$ we have

$$f(x) \leq f(y) + \nabla f(y)^T(x - y) + \frac{L}{2} \|x - y\|^2.$$

In-class assignment 3: Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be L -smooth over the set $X \subseteq \mathbb{R}^n$. Show that for any $x, y \in X$ we have

$$(\nabla f(x) - \nabla f(y))^T(x - y) \leq L \|x - y\|^2.$$

Lemma (Second derivative characterizations for quadratic function): Consider the quadratic function

$$f(x) \triangleq \frac{1}{2} x^T Q x + q^T x + d,$$

where Q is a symmetric matrix, $q \in \mathbb{R}^n$, and $d \in \mathbb{R}$. Then,

- (a) f is convex if and only if $Q \geq 0$.
- (b) f is μ -strongly convex if and only if $Q > 0$, where $\mu := \lambda_{\min}^Q$.
- (c) f is always L -smooth, where $L := \max \{|\lambda_{\min}^Q|, |\lambda_{\max}^Q|\}$.

In-class assignment 4:

- (a) Show part (c) of the above lemma.

(b) Extend the results of (a), (b), and (c) to cases where Q is not necessarily symmetric.

$$f(x_1, x_2) = x_1 x_2 = 0.5 x_1 x_2 + 0.5 x_2 x_1$$

We have that $\nabla f(x) = Qx + q$.

$$\|\nabla f(x) - \nabla f(y)\| = \|Qx + q - (Qy + q)\| = \|Q(x - y)\| \leq \|Q\| \|x - y\| = \max\{|\lambda_{\min}^Q|, |\lambda_{\max}^Q|\} \|x - y\|.$$

If Q is not symmetric, then we have the following:

$$Q \neq Q^T$$

$$x^T Qx = x^T Q^T x$$

$$x^T Qx = (x^T Qx)^T = x^T Q^T (x^T)^T = x^T Q^T x$$

$$\frac{1}{2} x^T Qx = \frac{1}{4} x^T Qx + \frac{1}{4} x^T Qx = \frac{1}{4} x^T Qx + \frac{1}{4} x^T Q^T x = \frac{1}{2} x^T \left(\frac{Q+Q^T}{2} \right) x$$

Let us define matrix $\bar{Q} = \frac{Q+Q^T}{2}$. Now we can write

$$f(x) = \frac{1}{2} x^T Qx + q^T x + d = \frac{1}{2} x^T \bar{Q}x + q^T x + d$$

we do have $\bar{Q} = \bar{Q}^T$. Thus, \bar{Q} is symmetric and so, all the results stated in the lemma hold for \bar{Q} . Thus, we always have

(a) f is convex if and only if $\frac{Q+Q^T}{2} \succeq 0$.

(b) f is μ -strongly convex if and only if $\frac{Q+Q^T}{2} \succ 0$, where $\mu := \lambda_{\min}^{\frac{Q+Q^T}{2}}$.

(c) f is always L -smooth, where $L := \max\left\{|\lambda_{\min}^{\frac{Q+Q^T}{2}}|, |\lambda_{\max}^{\frac{Q+Q^T}{2}}|\right\}$.

Lemma 1: Let $\|\cdot\|$ denote the Euclidean norm of a vector. For any $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$, the following holds.

(a) *Schwarz inequality:* $|u^T v| \leq \|u\| \|v\|$, with equality holding if and only if $u = \alpha v$ for some $\alpha \in \mathbb{R}$.

$$(b) \|u + v\|^2 = \|u\|^2 + 2u^T v + \|v\|^2.$$

$$(c) 2|u^T v| \leq \|u\|^2 + \|v\|^2.$$

$$(d) 2|u^T v| \leq \alpha \|u\|^2 + \frac{1}{\alpha} \|v\|^2 \text{ where } \alpha > 0 \text{ is a given scalar.}$$

$$(e) \|u + v\|^2 \leq 2\|u\|^2 + 2\|v\|^2.$$

Out-class assignment: Prove parts (b) to (e).

Gradient descent method

Let function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be given. We consider solving the following unconstrained optimization problem:

minimize $f(x)$ subject to: $x \in \mathbb{R}^n$.
--



Today, we will focus on the case where we choose a constant step-size.

Convergence and rate analysis

a) When $f(x)$ is an L -smooth and nonconvex function

b) When $f(x)$ is an L -smooth and convex function

c) When $f(x)$ is an L -smooth and μ -strongly convex function

The details are as follows:

a) The smooth nonconvex case

Theorem 1: Let $f^* := \min_{x \in \mathbb{R}^n} f(x)$ where f is L -smooth and possibly nonconvex. Then, when $\gamma \leq \frac{1}{L}$

$$\min_{k \in \{0, \dots, K-1\}} \|\nabla f(x_k)\|^2 \leq \frac{2\gamma^{-1}(f(x_0) - f^*)}{K} \quad \text{for all } K \geq 1.$$

Proof of Theorem 1:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2 && \leftarrow \text{from the Descent Lemma for } x := x_{k+1}, y := x_k \\ &= f(x_k) + \nabla f(x_k)^T (-\gamma \nabla f(x_k)) + \frac{L}{2} \|- \gamma \nabla f(x_k)\|^2 && \leftarrow \text{from the update rule } x_{k+1} := x_k - \gamma \nabla f(x_k) \\ &= f(x_k) - \gamma \|\nabla f(x_k)\|^2 + \frac{\gamma^2 L}{2} \|\nabla f(x_k)\|^2 && \leftarrow \text{from the definition of Euleadian norm sqare, i.e., } u^T u = \|u\|^2 \text{ for all } u \in \mathbb{R}^n \\ &= f(x_k) - \gamma \left(1 - \frac{\gamma L}{2}\right) \|\nabla f(x_k)\|^2 \\ &\leq f(x_k) - \frac{\gamma}{2} \|\nabla f(x_k)\|^2. && \leftarrow \text{obtained by assuming } \gamma \leq \frac{1}{L}, \text{ i.e., } 1 - \frac{\gamma L}{2} \geq \frac{1}{2} \end{aligned}$$

Thus, we have:

$$f(x_{k+1}) \leq f(x_k) - \frac{\gamma}{2} \|\nabla f(x_k)\|^2. \quad (1)$$

Observation 1: $f(x_k)$ is nonincreasing in terms of k .

Rearranging the terms, we obtain

$$\|\nabla f(x_k)\|^2 \leq \frac{2}{\gamma} (f(x_k) - f(x_{k+1})).$$

Summing both sides over $k = 0, \dots, K-1$ where $K \geq 1$, we obtain

$$\sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 \leq \frac{2}{\gamma} (f(x_0) - f(x_K)).$$

From $f^* \leq f(x_K)$, we have

$$\min_{k \in \{0, \dots, K-1\}} \|\nabla f(x_k)\|^2 \leq \frac{2\gamma^{-1}(f(x_0) - f^*)}{K}.$$

b) The smooth convex case

Theorem 2: Let $\min_{x \in \mathbb{R}^n} f(x)$ have at least one global optimal solution denoted by $x^* \in \mathbb{R}^n$, where f is L -smooth and convex. Then, when $\gamma \leq \frac{1}{L}$:

$$f(x_K) - f(x^*) \leq \frac{0.5\gamma^{-1}\|x_0 - x^*\|^2}{K} \quad \text{for all } K \geq 1.$$

Reading assignment: Read the proof of Theorem 2.

Proof of Theorem 2:

$$f(x_k) + \nabla f(x_k)^T (x^* - x_k) \leq f(x^*). \quad \leftarrow \text{from convexity inequality in the Lecture 20230120 for } y := x^*, x := x_k$$

Note that equation (1) in the proof of Theorem 1 holds here as well. Adding both sides of the preceding relation with equation (1), we obtain

$$f(x_{k+1}) + \nabla f(x_k)^T (x^* - x_k) \leq f(x^*) - \frac{\gamma}{2} \|\nabla f(x_k)\|^2$$

by rearranging terms $\Rightarrow f(x_{k+1}) - f(x^*) \leq -\nabla f(x_k)^T (x^* - x_k) - \frac{\gamma}{2} \|\nabla f(x_k)\|^2$

replacing $\nabla f(x_k)$ from the update rule $\Rightarrow f(x_{k+1}) - f(x^*) \leq \frac{1}{\gamma} (x_{k+1} - x_k)^T (x^* - x_k) - \frac{1}{2\gamma} \|x_{k+1} - x_k\|^2$

adding and subtracting x^* in the last term $\Rightarrow f(x_{k+1}) - f(x^*) \leq \frac{1}{\gamma} (x_{k+1} - x_k)^T (x^* - x_k) - \frac{1}{2\gamma} \|(x_{k+1} - x^*) - (x_k - x^*)\|^2$

$$\begin{aligned} \Rightarrow f(x_{k+1}) - f(x^*) &\leq \frac{1}{\gamma} (x_{k+1} - x_k)^T (x^* - x_k) \\ &\quad - \frac{1}{2\gamma} (\|x_{k+1} - x^*\|^2 + \|x_k - x^*\|^2 - 2(x_{k+1} - x^*)^T (x_k - x^*)) \end{aligned}$$

$$\Rightarrow f(x_{k+1}) - f(x^*) \leq \frac{1}{\gamma} (x_{k+1} - x_k)^T (x^* - x_k) - \frac{1}{2\gamma} \|x_{k+1} - x^*\|^2 + \frac{1}{2\gamma} \|x_k - x^*\|^2$$

Let us assume that f is L_0 -Lipschitz continuous.

$$\|f(x_k) - f(x^*)\| \leq L_0 \|x_k - x^*\|$$

$$\|f(x_k) - f(x^*)\|^2 \leq L_0^2 (1 - 2\gamma\mu + \gamma^2 L^2)^K \|x^* - x_0\|^2$$

c) The smooth strongly convex case

Theorem 3: Let $x^* \in \mathbb{R}^n$ denote the unique global optimal solution of $\min_{x \in \mathbb{R}^n} f(x)$, where f is L -smooth and μ -strongly convex. Then, when $\gamma < \frac{2\mu}{L^2}$:

$$\|x_K - x^*\|^2 \leq (1 - 2\gamma\mu + \gamma^2 L^2)^K \|x^* - x_0\|^2 \quad \text{for all } K \geq 1.$$

Proof of Theorem 3:

$$\|x_{k+1} - x^*\|^2 = \|x_k - \gamma \nabla f(x_k) - x^*\|^2 \quad \leftarrow \text{from the update rule for } x_{k+1}$$

$$= \|x_k - x^*\|^2 - 2\gamma \nabla f(x_k)^T (x_k - x^*) + \gamma^2 \|\nabla f(x_k)\|^2 \quad \leftarrow \text{from Lemma 1}$$

$$= \|x_k - x^*\|^2 - 2\gamma (\nabla f(x_k) - \nabla f(x^*))^T (x_k - x^*) + \gamma^2 \|\nabla f(x_k) - \nabla f(x^*)\|^2 \quad \leftarrow \text{why?}$$

$$\leq \|x_k - x^*\|^2 - 2\gamma\mu \|x_k - x^*\|^2 + \gamma^2 L^2 \|x_k - x^*\|^2 \quad \leftarrow \text{from in-class assignments 1 and 3}$$

$$= (1 - 2\gamma\mu + \gamma^2 L^2) \|x_k - x^*\|^2.$$

Let $\rho \triangleq 1 - 2\gamma\mu + \gamma^2 L^2$. Note that since $\gamma < \frac{2\mu}{L^2}$, we can show that $0 < \rho < 1$. Then, we obtain:

$$\|x_{k+1} - x^*\|^2 \leq \rho^{k+1} \|x_0 - x^*\|^2 \quad \text{for all } k \geq 0.$$

Implementation of the gradient method

Consider the regularized logistic regression loss function given as follows:

$$f(x) = \left(\frac{1}{N} \sum_{i=1}^N \ln(1 + \exp(-v_i u_i^T x)) \right) + \frac{\mu}{2} \|x\|^2,$$

where $\mu > 0$ is a known scalar and $u_i \in \mathbb{R}^n$ is the input vector associated with label i and $v_i \in \{-1, 1\}$ denotes the binary class of the i th label.

It can be shown that

$$\nabla f(x) = \left(\frac{1}{N} \sum_{i=1}^N \frac{-v_i u_i}{1 + \exp(v_i u_i^T x)} \right) + \mu x.$$

In-class assignment 5:

Suppose U is an ndarray of size $n \times N$ and v is an ndarray of size $1 \times N$. Write a line of code to return an ndarray whose i th column is equal to v_i multiplied by the i th column of U , for all i .

```
In [207]: 1 np.dot(A, np.diagflat(np.array([[2],[1]])))
```

```
Out[207]: array([[ 2,  2],
                [ 6,  4],
                [10,  6]])
```

Resolving a computational error in Python

```
In [208]: 1 from math import *
          2 exp(709)
```

```
Out[208]: 8.218407461554972e+307
```

Evaluate $\exp(710)$. What do you observe?

Resolving a computational error in Python

$\exp(709)$

$\exp(710)$

To avoid this overflow error, in the code, instead of writing

$$\exp(\text{some term}),$$

we use

$$\exp(\min(709, \text{that term})).$$

Theoretically, $\ln(e^x) = x$. Also, when x is large enough, $\ln(1 + e^x) = x$ is a very accurate approximation. So, in the code, instead of writing

$$\log(1 + \exp(\text{some term})),$$

some term

Define gradient and objective function

$$\nabla f(x) = \left(\frac{1}{N} \sum_{i=1}^N \frac{-v_i u_i}{1 + \exp(v_i u_i^T x)} \right) + \mu x, \quad f(x) = \left(\frac{1}{N} \sum_{i=1}^N \ln(1 + \exp(-v_i u_i^T x)) \right) + \frac{\mu}{2} \|x\|^2$$

Reading assignment: Read the remaining sections below.

Projection mapping

Projection onto a convex set is applied in a variety of constrained optimization solution methods.

Projection problem: Let $X \subseteq \mathbb{R}^n$ be a nonempty set and $\hat{x} \in \mathbb{R}^n$ be a given arbitrary vector. The projection problem is the problem of determining a point $x^* \in X$ that is the closest point to \hat{x} among all $x \in X$ with respect to the Euclidean distance. This problem is given by

$$\begin{array}{ll} \text{minimize} & \|x - \hat{x}\|_2^2 \\ \text{subject to} & x \in X. \end{array}$$

In general, this problem may not have an optimal solution and even when a solution exists, it may not be unique. However, when the set X is nonempty, closed and convex, the solution exists and it is unique. This is presented in the following result.

Projection Theorem:

Let set $X \subseteq \mathbb{R}^n$ be a nonempty closed convex set and \hat{x} be a given arbitrary vector.

(a) The projection problem has a unique optimal solution.

(b) A vector $x^* \in X$ is the solution to the projection problem if and only if

$$(x^* - \hat{x})^T (x - x^*) \geq 0 \quad \text{for all } x \in X.$$

Projection notation: Projection Theorem implies that the projection of a point onto a nonempty closed convex set is a unique point. We let $\mathcal{P}_X(\hat{x})$ denote the unique projection of a point \hat{x} onto a nonempty closed convex set X , i.e.,

$$\mathcal{P}_X(\hat{x}) = \operatorname{argmin}_{x \in X} \|x - \hat{x}\|_2^2.$$

The projection operator $\mathcal{P}_X(\cdot)$ has some important properties. Some of these properties are provided below.

Projection Properties: Let set $X \subseteq \mathbb{R}^n$ be a nonempty closed convex set. Consider the projection mapping $\mathcal{P}_X : \mathbb{R}^n \rightarrow X$.

(a) The projection mapping \mathcal{P}_X is nonexpansive, i.e.,

$$\|\mathcal{P}_X(x) - \mathcal{P}_X(y)\|_2 \leq \|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

(b) The distance function from a set, given by:

$$\text{dist}(x, X) = \|\mathcal{P}_X(x) - x\|_2,$$

is a convex function in terms of x .

Lemma (Properties of convex sets):

(a) For any collection $\{C_i \mid i \in I\}$ of convex sets, the set intersection $\cap_{i \in I} C_i$ is convex.

(b) If C is a convex set and $f : C \rightarrow \mathbb{R}$ is a convex function, the level sets $\{x \in C \mid f(x) \leq \alpha\}$ and $\{x \in C \mid f(x) < \alpha\}$ are convex for all $\alpha \in \mathbb{R}$.

Lemma (A variant of Jensen's inequality for convex functions): Consider a convex function $f : C \rightarrow \mathbb{R}$ over convex set C . Let $x_1, \dots, x_m \in C$ and $\alpha_1, \dots, \alpha_m \geq 0$ such that $\sum_{i=1}^m \alpha_i = 1$. Then, we have

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

Out-class assignment: Prove The above two lemmas.

Characterization of differentiable convex functions

When the function is convex and smooth, we have a nice set of properties given in the following proposition.

Proposition (First derivative characterizations): Let $C \subseteq \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable over \mathbb{R}^n .

(a) f is convex over C if and only if

$$f(y) + \nabla f(y)^T(x - y) \leq f(x), \quad \text{for all } x, y \in C.$$

(b) f is strictly convex over C if and only if the above inequality is strict whenever $x \neq y$.

(c) Let f be convex, For a scalar $L > 0$ the following statements are equivalent:

(c-i) $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, for all $x, y \in \mathbb{R}^n$.

(c-ii) $f(y) + \nabla f(y)^T(x - y) + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \leq f(x)$, for all $x, y \in \mathbb{R}^n$.

(c-iii) $f(x) \leq f(y) + \nabla f(y)^T(x - y) + \frac{L}{2}\|y - x\|^2$, for all $x, y \in \mathbb{R}^n$.

(c-iv) $(\nabla f(x) - \nabla f(y))^T(x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$, for all $x, y \in \mathbb{R}^n$.

(c-v) $(\nabla f(x) - \nabla f(y))^T(x - y) \leq L\|x - y\|^2$, for all $x, y \in \mathbb{R}^n$.

Characterization of convergence speed

Error function: To perform convergence analysis, we define an error function $e : \mathbb{R}^n \rightarrow \mathbb{R}$ to characterize the progress of the scheme at each iteration. A well-defined error function is the one that satisfies the following:

- For any $x \in \mathbb{R}^n$, we have $e(x) \geq 0$.
- $e(x^*) = 0$ where x^* is the local (global) minimum.

Examples of error functions include: $e(x) = \|x - x^*\|^2$, $e(x) = f(x) - f(x^*)$, and $e(x) = \|\nabla f(x)\|$.

Linear convergence:

We say $\{e(x_k)\}$ converges linearly if there exist some $\beta \in (0, 1)$ and q such that for all k :

$$e(x_k) \leq q\beta^k.$$

This is equivalent to finding a $\beta \in (0, 1)$ such that

$$\limsup_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} \leq \beta.$$

Stochastic Optimization (Module 3: SGD)

Stochastic gradient descent method: theory and implementation

Problem formulation

We consider solving the following [stochastic optimization](#) problem.

$$\begin{array}{ll} \text{minimize} & f(x) \triangleq \mathbb{E}[F(x, \xi)] \\ \text{subject to:} & \\ & x \in X. \end{array}$$

(P)

- Here, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is an unknown deterministic function.
- Here, $F : \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}$ is a known stochastic function.
- $\xi : \Omega \rightarrow \mathbb{R}^d$ is a d -dimensional random variable.
- $\mathbb{E}[\bullet]$ denotes the expectation operator with respect to ξ .

$$X = \{(x_1, x_2) \mid x_1 = 1, x_2 \leq 2\}$$

$$X = \{x \in \mathbb{R}^n \mid \|x\| \leq 1\}$$

$$X = \{x \in \mathbb{R}^n \mid Ax = b, x \geq 0\}$$

$$X = \{x \in \mathbb{R}^n \mid g(x) \leq 0\}$$

$$\hat{x} = [5, 0]$$

We can write for every given $x \in X$,

$$\mathbb{E}[F(x, \xi)] = \int_{\mathbb{R}^d} F(x, \xi) p(\xi) d\xi.$$

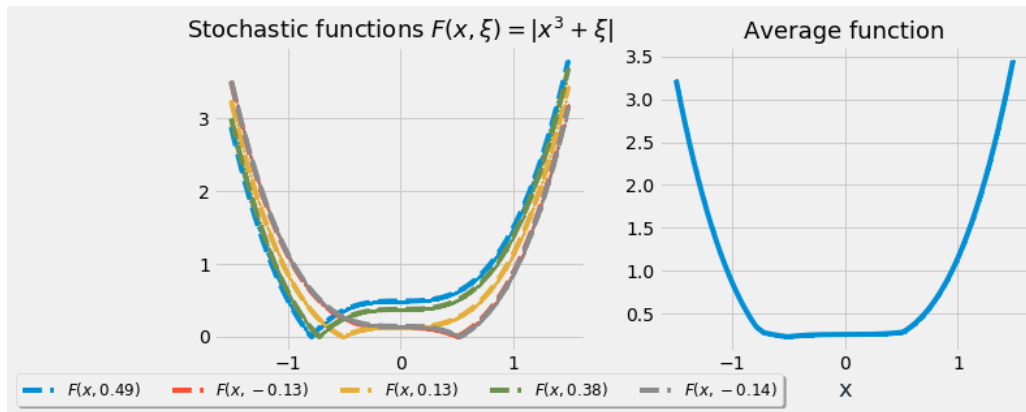
- **Challenge:** In implementing gradient method, computing $\nabla f(x)$ is either impossible or undesirable due to the presence of multi-dimensional integral involved in taking the expectation.

Example: Consider the function $f(x) := \mathbb{E}[x^3 + \xi]$ where $\xi \in \mathbb{R}$ is uniformly distributed in the interval $[-0.5, 0.5]$.

```

In [3]: 1 import matplotlib.pyplot as plt
2 import numpy as np
3 plt.style.use('fivethirtyeight')
4
5 x = np.arange(-1.5, 1.5, 0.01)
6 s = 5
7 z = 0*x
8
9 fig = plt.figure(figsize=(10,4))
10
11 plt.subplot(121)
12 for i in range(s):
13     ksi = 1*(np.random.rand()-0.5)
14     y = np.absolute(x**3+ksi)
15     #y = np.absolute(x**3-1)
16     plt.plot(x, y, marker='*', markersize=2, linestyle='dashed', label=r"$F(x, "+str(round(ksi,2))+")$", linewidth=4)
17     z+= y
18 plt.xlabel('x', color='#1C2833', fontsize=18)
19 plt.title(r"Stochastic functions $F(x, \xi) = |x^3 + \xi|$", fontsize=18)
20 plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.05),
21         fancybox=True, shadow=True, ncol=5, fontsize=12)
22
23 plt.subplot(122)
24 plt.plot(x, z/s, marker='*', markersize=2, linestyle='solid', label=r"$f(x)$", linewidth=4)
25
26 plt.xlabel('x', color='#1C2833', fontsize=18)
27 plt.title("Average function", fontsize=18)
28 plt.grid(True)

```



$$f(x) = \mathbb{E}[|x^3 + \xi|] = \int_{\mathbb{R}} |x^3 + \xi| p(\xi) d\xi = \int_{-0.5}^{0.5} |x^3 + \xi| d\xi.$$

Case 1: $x \geq \sqrt[3]{0.5}$. This implies $x^3 \geq 0.5$. Thus, $x^3 + \xi \geq 0$ for all $\xi \in [-0.5, 0.5]$.

Case 2: $x \leq -\sqrt[3]{0.5}$. This implies $x^3 \leq -0.5$. Thus, $x^3 + \xi \leq 0$.

Case 3: $x \in [-\sqrt[3]{0.5}, \sqrt[3]{0.5}]$.

In case 1:

$$f(x) = \int_{-0.5}^{0.5} (x^3 + \xi) d\xi = x^3 + \frac{\xi^2}{2} \Big|_{-0.5}^{0.5} = x^3.$$

In case 2:

$$f(x) = - \int_{-0.5}^{0.5} (x^3 + \xi) d\xi = -x^3 - \frac{\xi^2}{2} \Big|_{-0.5}^{0.5} = -x^3.$$

In-class assignment 1:

Find the formula of $f(x)$ in the third case.

In case 3:

$$\begin{aligned}
 f(x) &= \int_{-0.5}^{0.5} |x^3 + \xi| d\xi = \int_{-0.5}^{-x^3} |x^3 + \xi| d\xi + \int_{-x^3}^{0.5} |x^3 + \xi| d\xi \\
 &= - \int_{-0.5}^{-x^3} (x^3 + \xi) d\xi + \int_{-x^3}^{0.5} (x^3 + \xi) d\xi \\
 &= x^3(x^3 - 0.5) - \frac{\xi^2}{2} \Big|_{-0.5}^{-x^3} + x^3(x^3 + 0.5) + \frac{\xi^2}{2} \Big|_{-x^3}^{0.5} \\
 &= x^6 - 0.25
 \end{aligned}$$

Algorithm outline



Convergence and rate analysis

a) When $f(\bullet)$ is an L-smooth and nonconvex function and $X = \mathbb{R}^n$

b) When $f(\bullet)$ is a convex function and X is compact

Notation: We let the history of the method of random variables used up to iteration k be denoted by:

$$\mathcal{F}_k \triangleq \{x_0, \xi_0, \xi_1, \dots, \xi_{k-1}\} \quad \text{for all } k \geq 1,$$

and $\mathcal{F}_0 \triangleq \{x_0\}$. We also let $\mathbb{E}[\bullet \mid \mathcal{F}_k]$ denote the conditional expectation with respect to the filtration \mathcal{F}_k .

Assumption 1: We have:

i) $\mathbb{E}[\nabla F(x, \xi) \mid x] = \nabla f(x)$ for all $x \in X$.

ii) $\mathbb{E}[\|\nabla F(x, \xi) - \nabla f(x)\|^2 \mid x] \leq \sigma^2$ for all $x \in X$ for some $\sigma > 0$.

Assumption 1(i) implies that the stochastic gradient $\nabla F(\bullet, \xi)$ is an unbiased estimator of the true gradient of the objective function, that is $\nabla f(x)$.

Assumption 1(ii) implies that the stochastic gradient $\nabla F(\bullet, \xi)$ has a bounded variance.

Definition 1: Let us define the stochastic errors $w_k \triangleq \nabla F(x_k, \xi_k) - \nabla f(x_k)$ for all $k \geq 0$.

$$\nabla F(x_k, \xi_k) = \nabla f(x_k) + w_k$$

Remark: If \mathcal{F}_k is known, then x_k would be known and can be treated as a deterministic vector. Also, any x_t for $t \leq k$ would be known as well.

For this reason, in taking the conditional expectation $\mathbb{E}[\bullet \mid \mathcal{F}_k]$, we would treat x_k as a known value, but note that x_{k+1} must be treated a random variable. Because it requires ξ_k to be computed and $\xi_k \notin \mathcal{F}_k$.

What about w_k ? Is it deterministic when \mathcal{F}_k is known?

In-class assignment 2: Fill out the blanks.

$$\mathbb{E}[f(x_k) \mid \mathcal{F}_k] = f(x_k).$$

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \neq f(\mathbb{E}[x_{k+1} \mid \mathcal{F}_k]).$$

$$\mathbb{E}[\nabla f(x_k)^T w_k \mid \mathcal{F}_k] = \nabla f(x_k)^T \mathbb{E}[w_k \mid \mathcal{F}_k] = \nabla f(x_k)^T \mathbb{E}[\nabla F(x_k, \xi_k) - \nabla f(x_k) \mid \mathcal{F}_k] = \nabla f(x_k)^T (\mathbb{E}[\nabla F(x_k, \xi_k) \mid \mathcal{F}_k] - \nabla f(x_k)) = 0.$$

$$\mathbb{E}[\nabla F(x_k, \xi_k) \mid \mathcal{F}_k] = \nabla \mathbb{E}[F(x_k, \xi_k) \mid \mathcal{F}_k] = \nabla \mathbb{E}[f(x_k, \xi)] = \nabla f(x_k).$$

$$\mathbb{E}[\nabla F(x_k, \xi_{k-1}) \mid \mathcal{F}_k] = \nabla F(x_k, \xi_{k-1}).$$

$$\mathbb{E}[\nabla F(x_k, \xi_k)^T (x_{k+1} - x_k) \mid \mathcal{F}_k] \leq 0.5 \left(\mathbb{E}[\|\nabla F(x_k, \xi_k)\|^2 \mid \mathcal{F}_k] + \mathbb{E}[\|x_{k+1} - x_k\|^2 \mid \mathcal{F}_k] \right)$$

We assume that ξ_0, ξ_1, \dots are i.i.d. They all have the same distributions as ξ .

$$E(aX + b) = aE(X) + b$$

$$g_k^T w_k = \sum_{i=1}^n g_k^{(i)} w_k^{(i)}$$

$$\mathbb{E}[g_k^T w_k \mid \mathcal{F}_k] = \mathbb{E}[\sum_{i=1}^n g_k^{(i)} w_k^{(i)} \mid \mathcal{F}_k] = \sum_{i=1}^n \mathbb{E}[g_k^{(i)} w_k^{(i)} \mid \mathcal{F}_k] = \sum_{i=1}^n g_k^{(i)} \mathbb{E}[w_k^{(i)} \mid \mathcal{F}_k] = g_k^T \mathbb{E}[w_k \mid \mathcal{F}_k]$$

$$E(XY) \neq E(X)E(Y)$$

Lemma 1: Consider Definition 1. Under Assumption 1 we have for all $k \geq 0$:

$$\begin{aligned} \mathbb{E}[w_k \mid \mathcal{F}_k] &= \mathbf{0}_n \\ \mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] &\leq \sigma^2. \end{aligned}$$

Lemma 2: From the probability law, we have that $\mathbb{E}[\mathbb{E}[\bullet \mid \mathcal{F}_k]] = \mathbb{E}[\bullet]$.

$$E(X) = E_Y(E_X(X|Y))$$

a) The smooth nonconvex case

Theorem 1: Consider problem (P) and Algorithm 1. Let us define $f^* \triangleq \min_{x \in \mathbb{R}^n} f(x)$. Suppose $f^* > -\infty$. Let $X = \mathbb{R}^n$. Assume that f is L -smooth (and nonconvex). Then, the following results hold.

i) Let $\gamma_k \equiv \gamma := \frac{1}{\sqrt{K}}$. Then, we have:

$$\min_{k \in \{0, \dots, K-1\}} \mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{2(\mathbb{E}[f(x_0)] - f^*) + L\sigma^2}{\sqrt{K}} \quad \text{for all } K \geq L^2.$$

ii) Let γ_k be such that $\sum_{k=0}^{\infty} \gamma_k = \infty$ and $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$. Then, we have:

$$\lim_{k \rightarrow \infty} \mathbb{E}[\|\nabla f(x_k)\|^2] = 0.$$

In-class assignment 3: Read the proof and fill out the blanks.

Proof of Theorem 1 (i): From the Descent Lemma, for $y := x_{k+1}$ and $x := x_k$ we obtain:

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2.$$

Note that since $X = \mathbb{R}^n$, we have $x_{k+1} := x_k - \gamma_k \nabla F(x_k, \xi_k)$. From the update rule of the algorithm, we obtain:

$$f(x_{k+1}) \leq f(x_k) - \gamma_k \nabla f(x_k)^T (\nabla F(x_k, \xi_k)) + \frac{\gamma_k^2 L}{2} \|\nabla F(x_k, \xi_k)\|^2.$$

Invoking Definition 1, by replacing $\nabla F(x_k, \xi_k)$ for $\nabla f(x_k) + w_k$ we obtain

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \gamma_k \nabla f(x_k)^T (\nabla f(x_k) + w_k) + \frac{\gamma_k^2 L}{2} \|\nabla f(x_k) + w_k\|^2 \\ &= f(x_k) - \gamma_k \|\nabla f(x_k)\|^2 - \gamma_k \nabla f(x_k)^T w_k + \frac{L\gamma_k^2}{2} (\|\nabla f(x_k)\|^2 + \|w_k\|^2 + 2\nabla f(x_k)^T w_k). \end{aligned}$$

Taking conditional expectations on both sides, we obtain

$$\begin{aligned} \mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] &\leq \mathbb{E}[f(x_k) - \gamma_k \|\nabla f(x_k)\|^2 - \gamma_k \nabla f(x_k)^T w_k + \frac{L\gamma_k^2}{2} (\|\nabla f(x_k)\|^2 + \|w_k\|^2 + 2\nabla f(x_k)^T w_k) \mid \mathcal{F}_k] \\ &\leq f(x_k) - \gamma_k \|\nabla f(x_k)\|^2 - \gamma_k \nabla f(x_k)^T \mathbb{E}[w_k \mid \mathcal{F}_k] + \frac{L\gamma_k^2}{2} (\|\nabla f(x_k)\|^2 + \mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] + 2\mathbb{E}[\nabla f(x_k)^T w_k \mid \mathcal{F}_k]). \end{aligned}$$

From Lemma 1, we obtain

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \leq f(x_k) - \gamma_k \|\nabla f(x_k)\|^2 + \frac{L\gamma_k^2}{2} (\|\nabla f(x_k)\|^2 + \sigma^2).$$

Assuming $\gamma_k \leq \frac{1}{L}$, we obtain

$$\begin{aligned}\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] &\leq f(x_k) - \gamma_k \|\nabla f(x_k)\|^2 + \frac{\gamma_k}{2} (\|\nabla f(x_k)\|^2) + \frac{L\gamma_k^2}{2} (\sigma^2) \\ &\leq f(x_k) - \frac{\gamma_k}{2} (\|\nabla f(x_k)\|^2) + \frac{L\gamma_k^2 \sigma^2}{2}.\end{aligned}$$

Taking expectation with respect to \mathcal{F}_k from both sides and invoking Lemma 2, we obtain

$$\mathbb{E}[\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k]] \leq \mathbb{E}[f(x_k)] - \frac{\gamma_k}{2} \mathbb{E}[\|\nabla f(x_k)\|^2] + \frac{L\gamma_k^2 \sigma^2}{2}.$$

Thus, we have:

$$\boxed{\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] - \frac{\gamma_k}{2} \mathbb{E}[\|\nabla f(x_k)\|^2] + \frac{L\gamma_k^2 \sigma^2}{2}.} \quad (1)$$

Under a constant step-size $\gamma_k \equiv \gamma$ we obtain

$$\mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{2}{\gamma} (\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]) + L\gamma\sigma^2.$$

Summing both sides over $k = 0, \dots, K-1$ where $K \geq 1$, we obtain

$$\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{2}{\gamma} (\mathbb{E}[f(x_0)] - \mathbb{E}[f(x_K)]) + KL\gamma\sigma^2.$$

From $f^* \leq f(x_K)$ for any realization of x_K , we have:

$$\min_{k \in \{0, \dots, K-1\}} \mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{2(f(x_0) - f^*)}{K\gamma} + L\gamma\sigma^2.$$

The bound in part (i) is obtained by substituting $\gamma := \frac{1}{\sqrt{K}}$ above.

b) The convex case with a compact constraint set

Theorem 2: Consider problem (P) and Algorithm 1. Let Assumption 1 hold. Let $X \subset \mathbb{R}^n$. Suppose X is nonempty, compact, and convex. Assume that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and convex.

Let the step-size be diminishing given by:

$$\gamma_k := \frac{\gamma_0}{\sqrt{k+1}} \quad \text{for all } k \geq 0.$$

Consider the averaged iterate defined by:

$$\bar{x}_K \triangleq \frac{\sum_{k=0}^{K-1} x_k}{K} \quad \text{for all } K \geq 1.$$

Then, there exists some scalars $M > 0$ and $C > 0$ and an optimal solution vector $x^* \in X$ such that:

$$\mathbb{E}[f(\bar{x}_K)] - f(x^*) \leq \left(\frac{M}{2\gamma_0} + \gamma_0 (C^2 + \sigma^2) \right) \frac{1}{\sqrt{K}} \quad \text{for all } K \geq 2.$$

Proof of Theorem 2: From the update rule of the algorithm and the nonexpansivity of the projection, we obtain:

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq \|\mathcal{P}_X(x_k - \gamma_k \nabla f(x_k, \xi_k)) - \mathcal{P}_X(x^*)\|^2 \\ &\leq \|x_k - x^*\|^2 + \gamma_k^2 \|\nabla f(x_k, \xi_k)\|^2 - 2\gamma_k \nabla f(x_k, \xi_k)^T (x_k - x^*).\end{aligned}$$

From Definition 1 we obtain:

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 + \gamma_k^2 \|\nabla f(x_k) + w_k\|^2 - 2\gamma_k (\nabla f(x_k) + w_k)^T (x_k - x^*) \\ &= \|x_k - x^*\|^2 + \gamma_k^2 (\|\nabla f(x_k)\|^2 + \|w_k\|^2 + 2\nabla f(x_k)^T w_k) - 2\gamma_k (\nabla f(x_k) + w_k)^T (x_k - x^*).\end{aligned}$$

Taking conditional expectation on both sides, we obtain:

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^*\|^2 + \gamma_k^2 (\mathbb{E}[\|\nabla f(x_k)\|^2 \mid \mathcal{F}_k] + \mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] + 2\nabla f(x_k)^T \mathbb{E}[w_k \mid \mathcal{F}_k]) - 2\gamma_k (\nabla f(x_k) + \mathbb{E}[w_k \mid \mathcal{F}_k])^T (x_k - x^*).$$

Since X is compact and that f is continuous (due to convexity), there exists $C > 0$ such that $\sup_{x \in X} \|\nabla f(x)\| \leq C$. Invoking Assumption 1 and Lemma 1, we obtain:

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^*\|^2 + \gamma_k^2 (C^2 + \sigma^2) - 2\gamma_k \nabla f(x_k)^T (x_k - x^*).$$

From convexity of f , we have that $\nabla f(x_k)^T (x_k - x^*) \geq f(x_k) - f(x^*)$. We obtain:

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^*\|^2 + \gamma_k^2 (C^2 + \sigma^2) - 2\gamma_k (f(x_k) - f(x^*)).$$

Dividing both sides by $2\gamma_k$ and rearranging the terms, we obtain:

$$f(x_k) - f(x^*) \leq \frac{1}{2\gamma_k} (\|x_k - x^*\|^2 - \mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k]) + \frac{\gamma_k}{2} (C^2 + \sigma^2).$$

Taking expectation with respect to \mathcal{F}_k on both sides and invoking Lemma 2, we obtain:

$$\begin{aligned} \mathbb{E}[f(x_k)] - f(x^*) &\leq \frac{1}{2\gamma_k} (\mathbb{E}[\|x_k - x^*\|^2] - \mathbb{E}[\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k]]) + \frac{\gamma_k}{2} (C^2 + \sigma^2) \\ &= \frac{1}{2\gamma_k} (\mathbb{E}[\|x_k - x^*\|^2] - \mathbb{E}[\|x_{k+1} - x^*\|^2]) + \frac{\gamma_k}{2} (C^2 + \sigma^2). \end{aligned}$$

Adding and subtracting $\frac{1}{2\gamma_{k-1}} \mathbb{E}[\|x_k - x^*\|^2]$, we obtain:

$$\begin{aligned} \mathbb{E}[f(x_k)] - f(x^*) &\leq \left(\frac{1}{2\gamma_{k-1}} \mathbb{E}[\|x_k - x^*\|^2] - \frac{1}{2\gamma_k} \mathbb{E}[\|x_{k+1} - x^*\|^2] \right) \\ &\quad + \left(\frac{1}{2\gamma_k} - \frac{1}{2\gamma_{k-1}} \right) \mathbb{E}[\|x_k - x^*\|^2] + \frac{\gamma_k}{2} (C^2 + \sigma^2). \end{aligned}$$

From boundedness of the set X , there exists a scalar $M > 0$ such that $\max_{x \in X} \|x - x^*\|^2 \leq M$. Thus, we have:

$$\begin{aligned} \mathbb{E}[f(x_k)] - f(x^*) &\leq \left(\frac{1}{2\gamma_{k-1}} \mathbb{E}[\|x_k - x^*\|^2] - \frac{1}{2\gamma_k} \mathbb{E}[\|x_{k+1} - x^*\|^2] \right) \\ &\quad + \left(\frac{1}{2\gamma_k} - \frac{1}{2\gamma_{k-1}} \right) M + \frac{\gamma_k}{2} (C^2 + \sigma^2). \end{aligned}$$

Implementation of SGD method on MNIST

In some machine learning problems, the goal is to solve the following expected loss minimization:

$$\begin{aligned} &\text{minimize} && f(x) \triangleq \mathbb{E}[\mathcal{L}(u^T x, v)] \\ &\text{subject to:} && x \in \mathbb{R}^n, \end{aligned}$$

where \mathcal{L} is a loss function such as regularized logistic regression loss function and (u, v) is a random pair of input and output. To address this problem, one can assume a **sample average approximation (SAA)** as follows:

$$\begin{aligned} &\text{minimize} && f(x) \triangleq \frac{1}{|S_{\text{train}}|} \sum_{i \in S_{\text{train}}} \mathcal{L}(u_i^T x, v_i) \\ &\text{subject to:} && x \in \mathbb{R}^n. \end{aligned}$$

where $\mathcal{D} \triangleq \{(u_i, v_i) \in \mathbb{R}^n \times \{-1, +1\} \mid i \in S\}$ where $S \triangleq \{1, \dots, s\}$ denotes the index set and S is partitioned into S_{train} and S_{test} .

Consider the regularized logistic regression loss minimization problem given as follows:

$$\begin{aligned} &\text{minimize} && f(x) \triangleq \left(\frac{1}{|S_{\text{train}}|} \sum_{i \in S_{\text{train}}} \ln(1 + \exp(-v_i u_i^T x)) \right) + \frac{\mu}{2} \|x\|^2 \\ &\text{subject to:} && x \in \mathbb{R}^n. \end{aligned}$$

Applying the gradient descent method, we have:

$$x_{k+1} := x_k - \gamma \left(\left(\frac{1}{|S_{\text{train}}|} \sum_{i \in S_{\text{train}}} \frac{-v_i u_i}{1 + \exp(v_i u_i^T x_k)} \right) + \mu x_k \right). \quad (\text{Gradient Descent})$$

The challenge in implementing GD method is the expensive computation of the summation above when the training set has huge number of data points.

In SGD, we view (u, v) as ξ . So, each (u_i, v_i) chosen from the training set can be viewed as a realization (sample) of the random variable ξ .

$$x_{k+1} := x_k - \gamma \left(\frac{-v_{i^*} u_{i^*}}{1 + \exp(v_{i^*} u_{i^*}^T x_k)} + \mu x_k \right), \quad (\text{Stochastic Gradient Descent})$$

where i^* is randomly chosen from S_{train} using a uniform discrete distribution.

Binary classification

Suppose that a dataset $\{(u_i, v_i)\}_{i=1}^N$ is given where $u_i \in \mathbb{R}^n$ denotes certain characteristics (also called *features*) of an object i and $v_i \in \{-1, +1\}$ denotes the class of the object. If $v_i = 1$, object i belongs to a certain class and if $v_i = -1$ it does not.

The goal is to classify a new object using the given dataset. Using a linear predictor, we can consider a vector $x \in \mathbb{R}^n$ that represents the weight of the n features. We seek values of x such that for "most" of the given feature-labeled data (u_i, v_i) , we have

$$\begin{cases} x^T u_i > 0, & \text{if } v_i = +1, \\ x^T u_i < 0, & \text{if } v_i = -1. \end{cases}$$

We can view the term $v_i x^T u_i$ as a quantity whose sign characterizes misclassification (why?). Let us define a loss function that penalizes misclassification of the data. Then, it makes sense to solve the following optimization problem

$$\begin{aligned} \min_x \quad & \frac{1}{N} \sum_{i=1}^N \mathcal{L}(v_i u_i^T x), \quad (7) \\ \text{s.t.} \quad & x \in \mathbb{R}^n. \end{aligned}$$

Here $\mathcal{L} : \mathbb{R} \rightarrow \mathbb{R}$ denotes a function that penalizes negative values of its arguments. Some of the popular choices for \mathcal{L} are given as follows:

$$\begin{aligned} \mathcal{L}(z) &= \exp(-z), & (\text{exponential loss}) \\ \mathcal{L}(z) &= \ln(1 + \exp(-z)), & (\text{logistic loss}) \\ \mathcal{L}(z) &= \max\{0, 1 - z\}, & (\text{hinge loss}) \end{aligned}$$

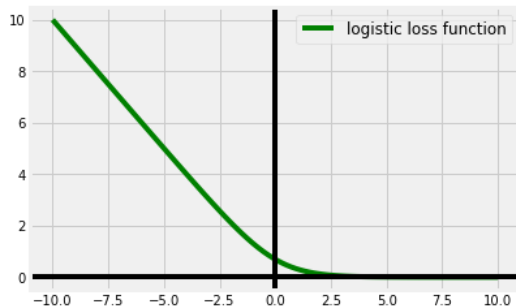
In the case that a logistic loss function is used, formulation (7) is called "logistic regression".

In the case that a hinge loss function is used, formulation (7) is called "support vector machines".

In-class assignment 4: Explain why the logistic loss function is a well-defined penalty function for misclassification.

Logistic loss function

```
In [18]: 1 import matplotlib.pyplot as plt
2 import numpy as np
3 from math import *
4
5 def logistic_loss_func(z):
6     return log(1+exp(-z))
7 x = np.linspace(-10, 10, 100)
8 y = np.asarray([logistic_loss_func(i) for i in x])
9 fig = plt.figure(figsize = (6, 4))
10 plt.plot(x,y,color='green',linestyle='solid',label="logistic loss function",linewidth=4)
11 plt.legend(loc="upper right",fontsize=12)
12 plt.axvline(x=0, color='k')
13 plt.axhline(y=0, color='k')
14 plt.grid(True)
15 plt.show()
```



How to classify a new sample after training the optimization model?

Let us denote the optimal solution of the preceding problem by x^* . To classify a new sample (i.e., not in the training data) denoted by $u_{\text{new}} \in \mathbb{R}^n$, it suffices to evaluate $u_{\text{new}}^T x^*$. Then,

$$\begin{cases} v_{\text{new}} = +1, & \text{if } u_{\text{new}}^T x^* > 0, \\ v_{\text{new}} = -1, & \text{if } u_{\text{new}}^T x^* < 0. \end{cases}$$

Projection onto a box

Let $X \subset \mathbb{R}^n$ be given as an n -dimensional box:

$$X = [l, u]^n,$$

where $l \leq u$ are scalars and $[l, u]^n$ denotes the Cartesian product of the one dimensional intervals $[l, u]$. It can be shown that:

(a) $\mathcal{P}_X(x) = \min\{\max\{l\mathbf{1}_n, x\}, u\mathbf{1}_n\}$, where $\mathbf{1}_n \in \mathbb{R}^n$ denotes a column vector with unit elements and the max and min operators are taken elementwise.

(b) $\mathcal{P}_X(x) = \min\{\max\{l\mathbf{1}_n, x\}, u\mathbf{1}_n\}$ if $x \in \mathbb{R}^n$ and $0 \leq x_i \leq u$ for all i .

Stochastic Optimization (Module 4: Quasi Newton methods)

Deterministic quasi-Newton methods: Motivation, theory, and implementation

Problem formulation

We consider solving the following [deterministic optimization](#) problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to:} & \\ & x \in \mathbb{R}^n. \end{array}$$

(P)

- Here, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is function.
- **Challenge:**
- Gradient descent method may perform poorly for ill-conditioned problems.
- Also, it does not utilize the curvature information of the objective function.

Recall that

$$\|x_K - x^*\|^2 \leq \rho^K \|x^* - x_0\|^2,$$

here $\rho := 1 - 2\gamma\mu + \gamma^2 L^2$. When $\rho \approx 1$, the method doesn't make much improvement.

For example, suppose $\rho = 0.9999$. Find the least number of iterations such that $\|x_K - x^*\|^2 \leq 0.001 \|x^* - x_0\|^2$.

$$\rho^K \leq 0.001 \quad \Leftrightarrow \quad K \ln(\rho) \leq \ln(0.001) \quad \Leftrightarrow \quad K \geq \frac{\ln(0.001)}{\ln(0.9999)} = 69,075.$$

The ratio $\frac{L}{\mu}$ is called the **condition number** of the problem. When it is very large (e.g., 10^{12}), the problem is **ill-conditioned**.

The issue with conditioning in gradient descent method

Consider minimizing

$$f(x_1, x_2) = \frac{a}{2}x_1^2 + \frac{b}{2}x_2^2 - x_1,$$

where $0 < a < b$ are given scalars.

In-class assignment 1: Find the condition number of function f in terms of a and b .

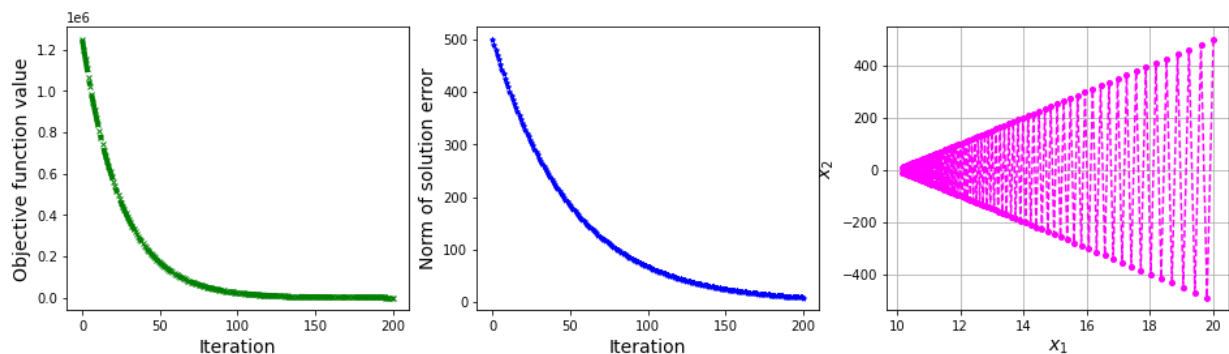
$$\frac{b}{a}$$

Suppose $b = 10$. Run the gradient method for $a \in \{0.1, 0.01, 0.001, 0.0001\}$.

```

In [13]: 1 import numpy as np
          2 import matplotlib.pyplot as plt
          3
          4 (a, b) = (10**-1, 10)
          5 opt_sol= np.array([[1/a], [0]])
          6 tol_a, tol_b = 2, 1
          7 Q = np.array([[a, 0], [0, b]])
          8 q = np.array([[1/a], [0]])
          9 x_0 = np.array([[20], [500]])
         10 k_max = 200
         11 stepsize = 2/(a+b)
         12
         13 def obj_f(x):
         14     output = 0.5*np.dot(x.T, np.dot(Q, x)) + np.dot(q.T, x)
         15     return output[0, 0]
         16
         17 def grad_f(x):
         18     return np.dot(Q, x) + q
         19
         20 def GD(x_0, k_max, stepsize):
         21     x_history = np.zeros((2, k_max+1))
         22     x_history[:, [0]] = x_0
         23     for k in range(k_max):
         24         x_history[:, [k+1]] = x_history[:, [k]] - stepsize*grad_f(x_history[:, [k]])
         25         #x_history[:, [k+1]] = x_history[:, [k]] - np.dot(np.array([[1/a, 0], [0, 1/b]]), grad_f(x_history[:, [k]]))
         26     return x_history
         27
         28 sol_history = GD(x_0, k_max, stepsize)
         29
         30 fvals = [obj_f(sol_history[:, [k]]) for k in range(k_max+1)]
         31 sol_dist = [np.linalg.norm(sol_history[:, [k]] - opt_sol) for k in range(k_max+1)]
         32
         33 fig = plt.figure(figsize=(16, 4))
         34 plt.subplot(131)
         35 plt.plot(range(k_max+1), fvals, color='green',
         36         marker='x', markersize=4, linestyle='dashdot', label="GD method", linewidth=4)
         37 plt.xlabel('Iteration', fontsize=14)
         38 plt.ylabel('Objective function value', fontsize=14)
         39
         40 plt.subplot(132)
         41 plt.plot(range(k_max+1), sol_dist, color='blue',
         42         marker='*', markersize=4, linestyle='dotted', linewidth=2)
         43 plt.xlabel('Iteration', fontsize=14)
         44 plt.ylabel('Norm of solution error', fontsize=14)
         45
         46 plt.subplot(133)
         47 plt.plot(sol_history[[0], :][0], sol_history[[1], :][0],
         48         marker='o', markersize=4, linestyle='dashed', color='magenta')
         49 plt.xlabel(r'$x_1$', fontsize=14)
         50 plt.ylabel(r'$x_2$', fontsize=14)
         51
         52 plt.grid(True)
         53 plt.show()

```



Newton's method

The update rule of the Newton's method is given as

$$x_{k+1} := x_k - \gamma_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k).$$

Newton method with $\gamma_k = 1$ and any arbitrary x_0 solves the above problem in only one iteration!

In-class assignment 2:

Revise the code above to implement the Newton's method. Use the constant stepsize $\gamma = 1$. Explain your observations.

$$c := \nabla f(x_k)$$

$$H := \nabla^2 f(x_k)$$

The question to find $d := H^{-1}c$

$$Hd = c$$

$$\min_d \quad 10$$

$$Hd = c.$$

$$\min_d \quad \frac{1}{2} \|Hd - c\|^2$$

$$d_{t+1} := d_t - \alpha_t H(Hd_t - c)$$

Quasi-Newton methods

Even though the Newton's method can enjoy a very fast convergence rate (a quadratic rate), for large scale problems, where n is large, it requires storage and computation of inverse of the Hessian. As such, in large-scale optimization, the Newton's method becomes inefficient due to **high computational cost and memory requirements** per iteration.

Motivated by these issues, and to improve the slow convergence speed of the gradient descent method, **quasi-Newton (QN) methods** were developed. This class of methods seeks to *approximate the Hessian without requiring to compute the Hessian or its inverse directly*. This way, the curvature information of the objective function is captured throughout the algorithm.

The class of quasi-Newton methods takes the form of

$$x_{k+1} := x_k + \gamma_k d_k, \quad \text{where } d_k := -D_k \nabla f(x_k).$$

Here $D_k > \mathbf{0}$ represents the Hessian inverse approximate at x_k and carries the curvature information of f .

Note that D_k is updated iteratively.

The intuition:

Let us consider the k th iteration. By quadratic approximation of f at x_k , we have

$$f(x_{k+1}) \approx f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{1}{2} (x_{k+1} - x_k)^T H_k (x_{k+1} - x_k),$$

where $H_k > \mathbf{0}$ is an approximation of Hessian. Let us assume $x_{k+1} := x_k + d_k$ with unit stepsizes. We have

$$f(x_{k+1}) \approx f(x_k) + \nabla f(x_k)^T d_k + \frac{1}{2} d_k^T H_k d_k.$$

Minimizing the right-hand side in terms of d_k , we obtain

$$\nabla f(x_k) + H_k d_k = 0 \quad \Rightarrow \quad d_k = -D_k \nabla f(x_k),$$

where $D_k \triangleq H_k^{-1}$.

In-class assignment 3: Under the condition that $D_k > \mathbf{0}$ for all $k \geq 0$, show that the above iterative method is a descent method.

How to update D_k ?

The main idea is to use the information of two successive iterations to approximate the Hessian inverse. Let us define:

$$\begin{aligned} s_k &\triangleq x_{k+1} - x_k, \\ y_k &\triangleq \nabla f(x_{k+1}) - \nabla f(x_k). \end{aligned}$$

Note that from the generalization of the first-order expansion for the gradient mapping we can write

$$y_k \approx \nabla^2 f(x_k) s_k.$$

This is in view of

$$\nabla f(x) \approx \nabla f(z) + \nabla^2 f(z)(x - z) \quad \Rightarrow \quad \nabla f(x) - \nabla f(z) \approx \nabla^2 f(z)(x - z)$$

For $x := x_{k+1}$ and $z := x_k$

Intuitive idea by William C. Davidon:

Consider the scheme

$$x_{k+1} := x_k + \alpha_k d_k, \quad \text{where } d_k := -D_k \nabla f(x_k).$$

Next, we develop an update rule for D_k .

Instead of computing D_k afresh every time, Davidon tried to build up **a recursive rule**. Suppose x_{k+1} is obtained and we wish to find H_{k+1} such that

$$f(x_{k+2}) \approx m_{k+1}(d) \triangleq f(x_{k+1}) + \nabla f(x_{k+1})^T d + \frac{1}{2} d^T H_{k+1} d.$$

Here $d \triangleq x - x_{k+1}$. We have

$$\nabla m_{k+1}(d) = \nabla f(x_{k+1}) + H_{k+1} d.$$

Question: What is the value of ∇m_{k+1} at x_{k+1} ?

The idea is to enforce $\nabla m_{k+1}(d)$ to match with $\nabla f(x_k)$ at x_k , i.e., at $d := x_k - x_{k+1}$. We obtain

$$\nabla m_{k+1}(d) = \nabla f(x_{k+1}) + H_{k+1}(x_k - x_{k+1}) = \nabla f(x_k),$$

which results in $\nabla f(x_{k+1}) - H_{k+1} s_k = \nabla f(x_k)$.

This implies that

$$y_k = H_{k+1} s_k \quad \text{or} \quad s_k = D_{k+1} y_k.$$

where we used $D_{k+1} \triangleq H_{k+1}^{-1}$ and the definition of s_k and y_k .

Secant equation: The equation $s_k = D_{k+1} y_k$ is called *secant equation*.

How to ensure a descent direction?

We want to find matrices D_k using $\{(s_i, y_i)\}_{i=0}^{k-1}$ such that $D_k > 0$ to ensure a descent direction at each iteration. The following lemma provides such condition.

Curvature condition: Note that in order to have a descent scheme, i.e., $D_k > 0$ for all k , we have to meet a condition called *curvature condition* given as follows:

$$s_k^T y_k > 0.$$

Lemma 1: (curvature condition) Let function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be strictly convex over \mathbb{R}^n . Show that for any k , the curvature condition is satisfied.

In-class assignment 4: Prove Lemma 1.

Hint: A function f is strictly convex if we have for all x, y

$$f(x) > f(y) + \nabla f(y)^T (x - y)$$

$$s_k = x_{k+1} - x_k$$

$$y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$$

$$f(x_{k+1}) > f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k)$$

$$f(x_k) > f(x_{k+1}) + \nabla f(x_{k+1})^T (x_k - x_{k+1})$$

$$f(x_{k+1}) + f(x_k) > f(x_k) + f(x_{k+1}) + \nabla f(x_k)^T (x_{k+1} - x_k) + \nabla f(x_{k+1})^T (x_k - x_{k+1})$$

$$0 > \nabla f(x_k)^T (x_{k+1} - x_k) - \nabla f(x_{k+1})^T (x_{k+1} - x_k)$$

$$0 > (\nabla f(x_k) - \nabla f(x_{k+1}))^T (x_{k+1} - x_k)$$

$$0 > -y_k^T s_k$$

$$y_k^T s_k > 0$$

BFGS method: The most popular quasi-Newton method

When the curvature condition is met, the linear system $s_k = D_{k+1} y_k$ may have infinite solutions for D_{k+1} . One way to find a good solution is to solve the following model:

$$\begin{aligned} \min_D \quad & \|D - D_k\|_F \\ \text{s.t.} \quad & Dy_k = s_k \\ & D > 0, \end{aligned}$$

where $\|\cdot\|_F$ is called Frobenius norm and is defined as

$$\|A\|_F \triangleq \sqrt{\sum_{i,j=1}^n a_{i,j}^2}.$$

Solving this problem will result in a recursive formula for D_{k+1} . In fact, it can be shown that D_{k+1} can be obtained from D_k and vectors s_k and y_k using the following recursive rule

$$D_{k+1} := (I - \rho_k s_k y_k^T) D_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T, \quad \rho_k \triangleq \frac{1}{y_k^T s_k}.$$

Lemma 2: $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be strictly convex over \mathbb{R}^n . Consider the quasi-Newton methods where D_k is generated by the BFGS rule. Then, $D_k > \mathbf{0}$ for any $k \geq 0$ and the method is descent at each iteration.

Pros and cons of BFGS method

The global convergence of the BFGS scheme can be established under strong convexity and smoothness of f . Under additional smoothness assumptions on the gradient map, i.e., Lipschitzian property of Hessian, it can be shown that the BFGS scheme attains a **superlinear** convergence rate.

However, at each iteration, in the BFGS update formula, D_k or alternatively, $\{(s_i, y_i)\}_{i=1}^k$ have to be stored.

$$f(x_{k+1}) \approx f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) = f(x_k) + \nabla f(x_k)^T (-\gamma D_k \nabla f(x_k)) = f(x_k) - \gamma \nabla f(x_k)^T D_k \nabla f(x_k)$$

```
In [1]: 1 import numpy as np
        2 n = 10**10
        3 np.random.normal(size=[n, n])

-----
ValueError                                Traceback (most recent call last)
/var/folders/vx/8xj88f1j3l9g2ll66w22zxx40000gn/T/ipykernel_79251/1211782645.py in <module>
      1 import numpy as np
      2 n = 10**10
----> 3 np.random.normal(size=[n, n])

mtrand.pyx in numpy.random.mtrand.RandomState.normal()

_common.pyx in numpy.random._common.cont()

ValueError: array is too big; `arr.size * arr.dtype.itemsize` is larger than the maximum possible size.
```

Limited memory BFGS (L-BFGS) method

To address this issue, a *limited-memory* variant of BFGS was developed in 1989 where at each iteration, $D_k \nabla f(x_k)$ is computed only by storing and manipulating a limited number of vectors, i.e., $\{(s_i, y_i)\}_{i=k-m+1}^k$.

It was shown that the L-BFGS method attains a **linear convergence** rate and performs significantly better in large scale optimization problems and specifically in the case where the condition number of the objective function is very large. This addresses a main shortcoming of the gradient descent method.

The underlying idea in limited memory BFGS is to do m updates of BFGS rule at each iteration and then restart the computation of Hessian inverse approximation at the next iteration.

The intuitive description of L-BFGS method:

Let $\{x_k\}$ be generated by the following quasi-Newton method:

$$x_{k+1} := x_k - \alpha \hat{D}_k \nabla f(x_k) \quad \text{with } \hat{D}_k \triangleq \begin{cases} \mathbf{I}, & \text{if } k < m, \\ D_{k,m}, & \text{if } k \geq m. \end{cases}$$

where matrix $D_{k,m}$ is recursively generated using the BFGS update rule m times as follows:

$$D_{k,j} := (I - \rho_i s_i y_i^T) D_{k,j-1} (I - \rho_i y_i s_i^T) + \rho_i s_i s_i^T, \quad \rho_i \triangleq \frac{1}{y_i^T s_i}, \quad \text{for all } j = 1, \dots, m, \quad (1)$$

where $i \triangleq k - (m - j)$ and $D_{k,0} \triangleq \frac{s_k^T y_k}{y_k^T y_k} I$ and we define

$$s_i \triangleq x_i - x_{i-1}, \\ y_i \triangleq \nabla f(x_i) - \nabla f(x_{i-1}).$$

In-class assignment 5:

a) What is the memory requirement of L-BFGS update rule?

$\mathcal{O}(mn)$

b) Let $m = 3$. Consider the $k = 10$ th iteration. List the pairs of (s_i, y_i) that are used to compute \hat{D}_{10} .

$j : 1, 2, 3$

$i : 8, 9, 10$

$(s_i, y_i) : (s_8, y_8), (s_9, y_9), (s_{10}, y_{10})$

b) For a any given m and $k \geq m$, list all the pairs of (s_i, y_i) that are used at iteration k to compute \hat{D}_k .

$j : 1, 2, \dots, m$

$$i : k - m + 1, \dots, k$$

$$(s_i, y_i) : \{(s_i, y_i)\}_{k-m+1}^k$$

An efficient implementation of L-BFGS method

In the following result, it is shown that the L-BFGS method does not require storing matrices and the direction of the method can be computed efficiently by a two-loop recursion with only $\mathcal{O}(mn)$ memory requirement.

Proposition (L-BFGS Two-loop recursion): Consider the L-BFGS scheme described above. Let vector r be generated at iteration k where $k \geq m$, as follows:

$$q := \nabla f(x_k)$$

for i **in** $(k, \dots, k - m + 1)$:

$$\rho_i := \frac{1}{y_i^T s_i}$$

$$\gamma_{k-i+1} := \rho_i s_i^T q$$

$$q := q - \gamma_{k-i+1} y_i$$

$$r := \frac{s_k^T y_k}{y_k^T y_k} q$$

for i **in** $(k - m + 1, \dots, k)$:

$$r := r + (\gamma_{k-i+1} - \rho_i y_i^T r) s_i$$

return r

Then, we have $r = \hat{D}_k \nabla f(x_k)$, where \hat{D}_k is the matrix of L-BFGS scheme given by (1).

Reading assignment: Read the following proof and do the outclass assignments.

Proof:

For clarity of the presentation, throughout this proof, q_{k-i+1} is used to denote the value of the vector $q \in \mathbb{R}^n$ after being updated at iteration k . Similarly, we use r_{i-k+m} to denote the value of the vector $r \in \mathbb{R}^n$ after being updated at iteration k . Also, we use the following definitions:

$$q_0 \triangleq \nabla f(x_k), \quad r_0 \triangleq \frac{s_k^T y_k}{y_k^T y_k} q_m,$$

$$\rho_i \triangleq \frac{1}{y_i^T s_i}, \text{ and } V_i \triangleq \mathbf{I} - \rho_i y_i s_i^T, \quad \text{for all } i = k - (m - 1), \dots, k.$$

Consider relation (1). By applying this recursive relation repeatedly, we obtain

$$\begin{aligned} D_{k,m} &= \left(\prod_{j=1}^m V_{k-(m-j)} \right)^T D_{k,0} \left(\prod_{j=1}^m V_{k-(m-j)} \right) \quad (2) \\ &+ \rho_{k-m+1} \left(\prod_{j=2}^m V_{k-(m-j)} \right)^T s_{k-m+1} s_{k-m+1}^T \left(\prod_{j=2}^m V_{k-(m-j)} \right) \\ &+ \rho_{k-m+2} \left(\prod_{j=3}^m V_{k-(m-j)} \right)^T s_{k-m+2} s_{k-m+2}^T \left(\prod_{j=3}^m V_{k-(m-j)} \right) \\ &+ \dots \\ &+ \rho_{k-1} V_k^T s_{k-1} s_{k-1}^T V_k \\ &+ \rho_k s_k s_k^T. \end{aligned}$$

Out-class assignment 1 (a): Use induction to prove the above equation for $D_{k,m}$.

Next, we derive a formula for q_i . We have

$$\begin{aligned} q_{k-i+1} &= q_{k-i} - \gamma_{k-i+1} y_i = q_{k-i} - \rho_i (s_i^T q_{k-i}) y_i = q_{k-i} - \rho_i (y_i s_i^T) q_{k-i} \\ &= (\mathbf{I} - \rho_i y_i s_i^T) q_{k-i} = V_i q_{k-i}, \quad \text{for all } i = k, k-1, \dots, k-m+1. \end{aligned}$$

From the preceding relation, we obtain

$$q_\ell = \left(\prod_{j=m-\ell+1}^m V_{k-(m-j)} \right) q_0, \quad \text{for all } \ell = 1, 2, \dots, m. \quad (3)$$

From the update rule for γ_{k-i+1} in the two-loop recursion scheme, using the definition of ρ_i , and applying the previous relation, we have $\gamma_1 = \rho_k s_k^T q_0$ and

$$\gamma_\ell = \rho_{k-\ell+1} s_{k-\ell+1}^T \left(\prod_{j=m-\ell+2}^m V_{k-(m-j)} \right) q_0, \quad \text{for all } \ell = 2, 3, \dots, m. \quad (4)$$

Multiplying both sides of (2) by q_0 and employing (3) and (4), we obtain

$$\begin{aligned} D_{k,m} q_0 &= \left(\prod_{j=1}^m V_{k-(m-j)} \right)^T D_{k,0} q_m + \left(\prod_{j=2}^m V_{k-(m-j)} \right)^T s_{k-m+1} \gamma_m \\ &\quad + \left(\prod_{j=3}^m V_{k-(m-j)} \right)^T s_{k-m+2} \gamma_{m-1} + \dots + V_k^T s_{k-1} \gamma_2 + s_k \gamma_1. \end{aligned} \quad (5)$$

Next, we derive a formula for r_i . From the two-loop recursion scheme, we have

$$\begin{aligned} r_{i-k+m} &= r_{i-k+m-1} + (\gamma_{k-i+1} - \rho_i y_i^T r_{i-k+m-1}) s_i \\ &= r_{i-k+m-1} - \rho_i s_i y_i^T r_{i-k+m-1} + \gamma_{k-i+1} s_i \\ &= V_i^T r_{i-k+m-1} + \gamma_{k-i+1} s_i, \quad \text{for all } i = k-m+1, \dots, k-1, k. \end{aligned}$$

Combining the preceding two relations, we obtain

$$r_\ell = V_{k-(m-\ell)}^T r_{\ell-1} + \gamma_{m-\ell+1} s_{k-(m-\ell)}, \quad \text{for all } \ell = 1, 2, \dots, m.$$

Using the preceding equation repeatedly, we obtain

$$\begin{aligned} r_m &= \left(\prod_{j=1}^m V_{k-(m-j)} \right)^T r_0 + \gamma_m \left(\prod_{j=2}^m V_{k-(m-j)} \right)^T s_{k-m+1} \\ &\quad + \gamma_{m-1} \left(\prod_{j=3}^m V_{k-(m-j)} \right)^T s_{k-m+2} + \dots + \gamma_2 V_k^T s_{k-1} + \gamma_1 s_k. \end{aligned} \quad (6)$$

Out-class assignment 1 (b): Use induction to prove the above equation for r_m .

From (5) and (6), and the definition of r_0 and q_0 , we obtain $r_m = D_{k,m} q_0 = \hat{D}_k \nabla f(x_k)$.

Suppose $a, b, c \in \mathbb{R}^n$. We have

$$(ab^T)c = (b^T c)a.$$

Stochastic quasi-Newton methods: theory and implementation

Problem formulation

We consider solving the following [stochastic optimization](#) problem.

$$\begin{aligned} &\text{minimize} && f(x) \triangleq \mathbb{E}[F(x, \xi)] \\ &\text{subject to:} && \\ &&& x \in \mathbb{R}^n. \end{aligned}$$

(P)

- Here, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is an unknown deterministic function.
- Here, $F : \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}$ is a known stochastic function.
- $\xi : \Omega \rightarrow \mathbb{R}^d$ is a d -dimensional random variable.
- $\mathbb{E}[\bullet]$ denotes the expectation operator with respect to ξ .

Convergence and rate analysis

a) When $F(\bullet, \xi)$ is an L-smooth and nonconvex function

b) When $F(\bullet, \xi)$ is a strongly convex function

Notation: We let the history of the method of random variables used up to iteration k be denoted by:

$$\mathcal{F}_k \triangleq \{x_0, \xi_0, \xi_1, \dots, \xi_{k-1}\} \quad \text{for all } k \geq 1,$$

and $\mathcal{F}_0 \triangleq \{x_0\}$. We also let $\mathbb{E}[\bullet \mid \mathcal{F}_k]$ denote the conditional expectation with respect to the filtration \mathcal{F}_k .

Assumption 1: We have:

- i) $\mathbb{E}[\nabla F(x, \xi) \mid x] = \nabla f(x)$ for all $x \in X$.
- ii) $\mathbb{E}[\|\nabla F(x, \xi) - \nabla f(x)\|^2 \mid x] \leq \sigma^2$ for all $x \in X$ for some $\sigma > 0$.

Definition 1: Let us define the stochastic errors $w_k \triangleq \nabla F(x_k, \xi_k) - \nabla f(x_k)$ for all $k \geq 0$.

Lemma 1: Consider Definition 1. Under Assumption 1 we have for all $k \geq 0$:

$$\begin{aligned}\mathbb{E}[w_k \mid \mathcal{F}_k] &= \mathbf{0}_n \\ \mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] &\leq \sigma^2.\end{aligned}$$

Lemma 2: From the probability law, we have that $\mathbb{E}[\mathbb{E}[\bullet \mid \mathcal{F}_k]] = \mathbb{E}[\bullet]$.

Assumption 2: Let the following hold for all $k \geq 0$:

- a) The matrix $H_k \in \mathbb{R}^{n \times n}$ is \mathcal{F}_k -measurable, i.e., $\mathbb{E}[H_k \mid \mathcal{F}_k] = H_k$.
- b) Matrix H_k is symmetric and positive definite, i.e., there exist positive scalars $\lambda_{\min}, \lambda_{\max}$ we have

$$\lambda_{\min} \mathbf{I} \leq H_k \leq \lambda_{\max} \mathbf{I}, \quad \text{for all } k \geq 0.$$

$A \geq 0_{n \times n}$ means that A is positive semi-definite

$A \geq B$ means that $A - B \geq 0_{n \times n}$

$A \geq \eta \mathbf{I}_n$ mean that the minimum eigenvalue of A is greater or equal to η .

Outline of the SQN method

Let x_k be generated recursively as follows for $k \geq 0$:

$$x_{k+1} = x_k - \gamma_k H_k \nabla F(x_k, \xi_k).$$

Rate analysis for the smooth nonconvex case

Theorem 1: Consider problem (P) and the SQN scheme. Let us define $f^* \triangleq \min_{x \in \mathbb{R}^n} f(x)$ where $f(x) = \mathbb{E}[F(x, \xi)]$. Suppose Assumptions 1 and 2 hold. Let $X = \mathbb{R}^n$ and let f be L -smooth (and possibly nonconvex). Then, the following results hold.

- i) Let $\gamma_k \equiv \gamma := \frac{1}{\sqrt{K}}$. Then, we have:

$$\min_{k \in \{0, \dots, K-1\}} \mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{2(\mathbb{E}[f(x_0)] - f^*) + L\lambda_{\max}^2 \sigma^2}{\lambda_{\min} \sqrt{K}} \quad \text{for all } K \geq \frac{L^2 \lambda_{\max}^4}{\lambda_{\min}^2}.$$

- ii) Let γ_k be such that $\sum_{k=0}^{\infty} \gamma_k = \infty$ and $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$. Then, we have:

$$\lim_{k \rightarrow \infty} \mathbb{E}[\|\nabla f(x_k)\|^2] = 0.$$

In-class assignment 6: Read the proof and fill out the blanks.

Proof of Theorem 1:

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2.$$

Note that we have $x_{k+1} := x_k - \gamma_k H_k \nabla F(x_k, \xi_k)$. From the update rule of the algorithm, we obtain

$$f(x_{k+1}) \leq f(x_k) - \gamma_k \nabla f(x_k)^T H_k \nabla F(x_k, \xi_k) + \frac{\gamma_k^2 L}{2} \|H_k \nabla F(x_k, \xi_k)\|^2.$$

Invoking Definition 1, by replacing $\nabla F(x_k, \xi_k)$ for $\nabla f(x_k) + w_k$ we obtain

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) - \gamma_k \nabla f(x_k)^T H_k (\nabla f(x_k) + w_k) + \frac{\gamma_k^2 L}{2} \|\nabla f(x_k) + w_k\|^2 \\
&= f(x_k) - \gamma_k \nabla f(x_k)^T H_k \nabla f(x_k) - \gamma_k \nabla f(x_k)^T H_k w_k + \frac{\gamma_k^2 L}{2} (\nabla f(x_k) + w_k)^T H_k (\nabla f(x_k) + w_k).
\end{aligned}$$

Note that from Proposition A.2 from the Lecture notes on Jan. 20, we can write

$$\nabla f(x_k)^T H_k \nabla f(x_k) \geq \lambda_{\min} \|\nabla f(x_k)\|^2.$$

Invoking Propositions A.1 and A.2 from the Lecture notes on Jan. 20, we have

$$(\nabla f(x_k) + w_k)^T H_k (\nabla f(x_k) + w_k) \leq \lambda_{\max}^2 \|\nabla f(x_k) + w_k\|^2.$$

Taking conditional expectations on both sides, we obtain

$$\begin{aligned}
\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] &\leq \mathbb{E} \left[f(x_k) - \lambda_{\min} \gamma_k \|\nabla f(x_k)\|^2 - \gamma_k \nabla f(x_k)^T H_k w_k + \frac{L \lambda_{\max}^2 \gamma_k^2}{2} (\|\nabla f(x_k)\|^2 + \|w_k\|^2 + 2 \nabla f(x_k)^T w_k) \mid \mathcal{F}_k \right] \\
&\leq f(x_k) - \lambda_{\min} \gamma_k \|\nabla f(x_k)\|^2 - \gamma_k \nabla f(x_k)^T H_k \mathbb{E}[w_k \mid \mathcal{F}_k] + \frac{L \lambda_{\max}^2 \gamma_k^2}{2} (\|\nabla f(x_k)\|^2 + \mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] + 2 \mathbb{E}[\nabla f(x_k)^T w_k \mid \mathcal{F}_k]).
\end{aligned}$$

From Lemma 1, we obtain

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \leq f(x_k) - \lambda_{\min} \gamma_k \|\nabla f(x_k)\|^2 + \frac{L \lambda_{\max}^2 \gamma_k^2}{2} (\|\nabla f(x_k)\|^2 + \sigma^2).$$

Assuming $\gamma_k \leq \frac{\lambda_{\min}}{L \lambda_{\max}^2}$, we obtain

$$\begin{aligned}
\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] &\leq f(x_k) - \lambda_{\min} \gamma_k \|\nabla f(x_k)\|^2 + \frac{\lambda_{\min} \gamma_k}{2} \|\nabla f(x_k)\|^2 + \frac{L \gamma_k^2 \lambda_{\max}^2}{2} (\sigma^2) \\
&\leq f(x_k) - \frac{\lambda_{\min} \gamma_k}{2} \|\nabla f(x_k)\|^2 + \frac{L \lambda_{\max}^2 \gamma_k^2 \sigma^2}{2}.
\end{aligned}$$

Taking expectation with respect to \mathcal{F}_k from both sides and invoking the total probability rule, we obtain

$$\mathbb{E}[\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k]] \leq \mathbb{E}[f(x_k)] - \frac{\lambda_{\min} \gamma_k}{2} \mathbb{E}[\|\nabla f(x_k)\|^2] + \frac{L \lambda_{\max}^2 \gamma_k^2 \sigma^2}{2}.$$

Thus, we have:

$$\boxed{\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] - \frac{\lambda_{\min} \gamma_k}{2} \mathbb{E}[\|\nabla f(x_k)\|^2] + \frac{L \lambda_{\max}^2 \gamma_k^2 \sigma^2}{2}.} \quad (1)$$

Under a constant step-size $\gamma_k \equiv \gamma$ we obtain

$$\mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{2}{\lambda_{\min} \gamma} (\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]) + \frac{L \lambda_{\max}^2 \gamma \sigma^2}{\lambda_{\min}}.$$

Summing both sides over $k = 0, \dots, K-1$ where $K \geq 1$, we obtain

$$\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{2}{\lambda_{\min} \gamma} (\mathbb{E}[f(x_0)] - \mathbb{E}[f(x_K)]) + K \frac{L \lambda_{\max}^2 \gamma \sigma^2}{\lambda_{\min}}.$$

From $f^* \leq f(x_K)$ for any realization of x_K , we have:

$$\min_{k \in \{0, \dots, K-1\}} \mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{2(f(x_0) - f^*)}{K \lambda_{\min} \gamma} + \frac{L \lambda_{\max}^2 \gamma \sigma^2}{\lambda_{\min}}.$$

The bound in part (i) is obtained by substituting $\gamma := \frac{1}{\sqrt{K}}$ above.

Appendix

Definition (little o notation):

For a positive integer p and a map $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we write $h(x) = o(\|x\|^p)$ if

$$\lim_{k \rightarrow \infty} \frac{h(x_k)}{\|x_k\|^p} = 0$$

for all (nonzero) sequences $\{x_k\}$ such that $x_k \rightarrow 0$.

Remark: Loosely speaking, the above notation means $o(\|x\|^p)$ becomes much smaller than $\|x\|^p$, when $\|x\|$ is small enough.

Example: $h(x) = x^3$ is $o(x)$.

Proposition (Second order expansion): Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable over an open sphere S , and let $x \in S$. Then for all y such that $x + y \in S$, we have:

- There exists an $\alpha \in [0, 1]$ such that

$$f(x + y) = f(x) + y^T \nabla f(x) + \frac{1}{2} y^T \nabla^2 f(x + \alpha y) y.$$

- The following holds:

$$f(x + y) = f(x) + y^T \nabla f(x) + \frac{1}{2} y^T \nabla^2 f(x) y + o(\|y\|^2).$$

In-class assignment 7:

1. Generate an ndarray of size 20 by 1 of random integer numbers ranging from -10 to 50.
2. Create a dataframe called "df_data" whose only column is the ndarray with the name "Some data".
3. Randomly shuffle the indices of the dataframe using np.random.permutation. Call it "shuffled_indices".
4. Create a new array that is the first 15 elements of shuffled indices. Call this train_indices.
5. Use iloc[] command to generate the subset of the df_data that is corresponding to the train_indices. Call this dataframe "df_train".
6. Create a dictionary called dict_train and store the df_train as an item in this dictionary.
7. Repeat steps 4 and 5 for the other 5 elements for test data.

Stochastic Optimization (Module 5: Zeroth-Order SGD)

Zeroth-order methods are some of the most important iterative methods in convex and nonconvex optimization.

In minimizing a function f over some set X , there is one main reason for the need to a **zeroth-order method**: The gradient mapping of the objective function may not be available.

This could be because of the following reasons:

- The objective function is nondifferentiable.
- The gradient's closed-form formula is not available. Often in this case, we may not even know if f is differentiable or not. Even if it is, we may not know what ∇f is to use it in gradient-type methods (a.k.a. first-order methods).

Spherical randomized smoothing

Let function $h : \mathbb{R}^n \rightarrow \mathbb{R}$ be given. Define the function h_η as

$$h_\eta(x) \triangleq \mathbb{E}_{u \in \mathbb{B}}[h(x + \eta u)],$$

where $\eta > 0$ is an arbitrary smoothing parameter, u is a random vector in the unit ball \mathbb{B} , defined as

$$\mathbb{B} \triangleq \{u \in \mathbb{R}^n \mid \|u\| \leq 1\}.$$

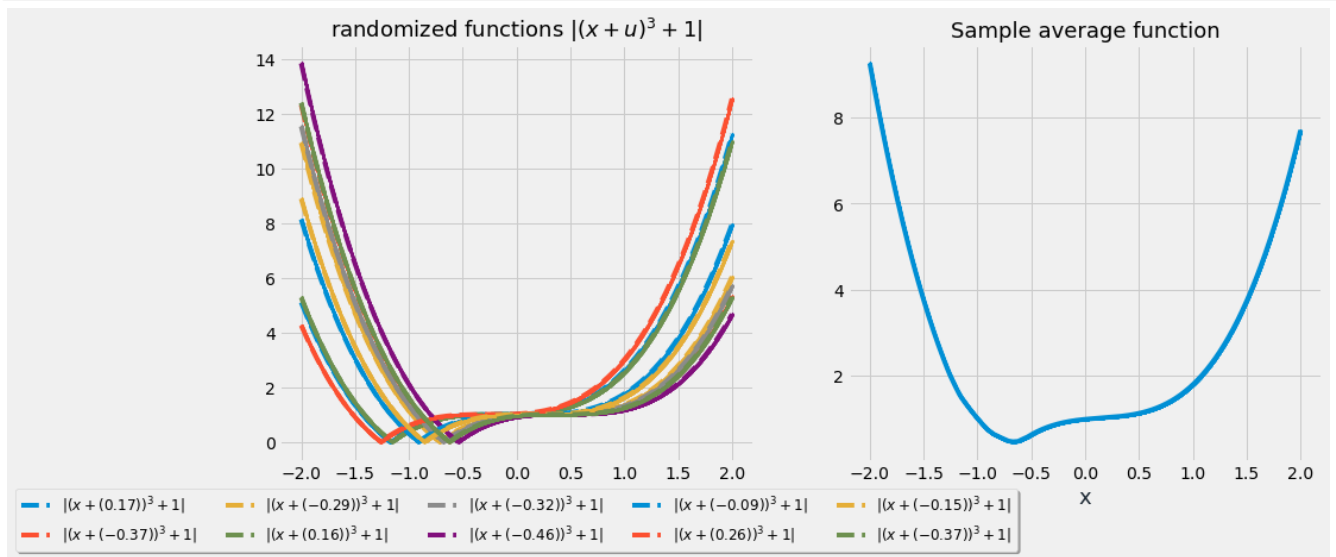
Example: Consider the function $f(x) := |x^3 + 1|$.

Define $f_{0.5}(x) := \mathbb{E}[|(x + 0.5u)^3 + 1|]$ where $u \in \mathbb{R}$ is uniformly distributed in the interval $[-1, 1]$.


```

In [5]: 1 import matplotlib.pyplot as plt
2 import numpy as np
3 plt.style.use('fivethirtyeight')
4
5 x = np.arange(-2, 2, 0.001)
6 s = 10
7 z = 0*x
8
9 fig = plt.figure(figsize=(14,6))
10
11 plt.subplot(121)
12 for i in range(s):
13     u = 1*(np.random.rand()-0.5)
14     y = np.absolute((x+u)**3+1)
15     plt.plot(x, y, marker='*', markersize=2, linestyle='dashed', label=r"$|(x+(\text{round}(u,2))+))^3+1|$", linewidth=4)
16     z += y
17 plt.xlabel('x', color='#1C2833', fontsize=18)
18 plt.title(r"randomized functions $|(x+u)^3+1|$", fontsize=18)
19 plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.05),
20         fancybox=True, shadow=True, ncol=5, fontsize=12)
21
22 plt.subplot(122)
23 plt.plot(x, z/s, marker='*', markersize=2, linestyle='solid', label=r"$f(x)$", linewidth=4)
24
25 plt.xlabel('x', color='#1C2833', fontsize=18)
26 plt.title("Sample average function", fontsize=18)
27 plt.grid(True)

```



Remark

Let us define $\tilde{u} := \eta u$. Note that \tilde{u} is uniformly distributed over $\eta\mathbb{B}$. In view of $u = \frac{1}{\eta}\tilde{u}$ (viewing u as a function of \tilde{u}) we have

$$\mathbb{E}_{u \in \mathbb{B}} [h(x + \eta u)] = \int h(x + \eta u) p(\tilde{u}) d\tilde{u} = \int_{\eta\mathbb{B}} h\left(x + \eta\left(\frac{1}{\eta}\tilde{u}\right)\right) p(\tilde{u}) d\tilde{u} = \int_{\eta\mathbb{B}} h(x + \tilde{u}) p(\tilde{u}) d\tilde{u} = \mathbb{E}_{\tilde{u} \in \eta\mathbb{B}} [h(x + \tilde{u})].$$

In view of this relation, throughout, we may use $\mathbb{E}_{u \in \mathbb{B}} [h(x + \eta u)]$ and $\mathbb{E}_{\tilde{u} \in \eta\mathbb{B}} [h(x + \tilde{u})]$ interchangeably.

$$f_{0.5}(x) = \mathbb{E}_{u \in \mathbb{B}} [|x + 0.5u|^3 + 1] = \mathbb{E}_{\tilde{u} \in 0.5\mathbb{B}} [h(x + \tilde{u})] = \int_{-0.5}^{0.5} |(x + \tilde{u})^3 + 1| d\tilde{u}.$$

Case 1: $x \geq -0.5$. Thus, $(x + \tilde{u})^3 + 1 \geq 0$ for all $\tilde{u} \in [-0.5, 0.5]$.

Case 2: $x \leq -1.5$. Thus, $(x + \tilde{u})^3 + 1 \leq 0$ for all $\tilde{u} \in [-0.5, 0.5]$.

Case 3: $x \in [-1.5, -0.5]$.

In case 1:

$$\begin{aligned}
f_{0.5}(x) &= \int_{-0.5}^{0.5} ((x + \tilde{u})^3 + 1) d\tilde{u} = 1 + \frac{1}{4}(x + \tilde{u})^4 \Big|_{-0.5}^{0.5} = 1 + \frac{1}{4}(x + 0.5)^4 - \frac{1}{4}(x - 0.5)^4 \\
&= 1 + \frac{1}{4} \sum_{i=0}^4 \binom{4}{i} x^{4-i} 0.5^i - \frac{1}{4} \sum_{i=1}^4 \binom{4}{i} x^{4-i} (-0.5)^i \\
&= 1 + \frac{1}{4} \sum_{i=0}^4 \binom{4}{i} x^{4-i} (0.5^i - (-0.5)^i) \\
&= 1 + \frac{1}{4} \sum_{i=1,3}^4 \binom{4}{i} x^{4-i} (0.5^i - (-0.5)^i) \\
&= x^3 + 0.25x + 1.
\end{aligned}$$

In case 2:

$$f_{0.5}(x) = - \int_{-0.5}^{0.5} ((x + \tilde{u})^3 + 1) d\tilde{u} = -\frac{1}{4}(x + \tilde{u})^4 \Big|_{-0.5}^{0.5} = -\frac{1}{4}(x + 0.5)^4 + \frac{1}{4}(x - 0.5)^4 = -x^3 - 0.25x - 1.$$

In-class assignment 1:

Find the formula of $f(x)$ in the third case.

$$\begin{aligned}
f_{0.5}(x) &= \int_{-0.5}^{0.5} |(x + \tilde{u})^3 + 1| d\tilde{u} = \int_{-0.5}^{-x-1} |(x + \tilde{u})^3 + 1| d\tilde{u} + \int_{-x-1}^{0.5} |(x + \tilde{u})^3 + 1| d\tilde{u} \\
&= - \int_{-0.5}^{-x-1} ((x + \tilde{u})^3 + 1) d\tilde{u} + \int_{-x-1}^{0.5} ((x + \tilde{u})^3 + 1) d\tilde{u} \\
&= (2x + 2) - \left(\frac{1}{4}(x + \tilde{u})^4 \Big|_{-0.5}^{-x-1} \right) + \frac{1}{4}(x + \tilde{u})^4 \Big|_{-x-1}^{0.5} \\
&= (2x + 2) - \left(\frac{1}{4} - \frac{1}{4}(x - 0.5)^4 \right) + \frac{1}{4}(x + 0.5)^4 - \frac{1}{4} \\
&= (2x + 2) + \frac{1}{4}(x + 0.5)^4 + \frac{1}{4}(x - 0.5)^4 - \frac{1}{2} \\
&= (2x + 2) + \frac{1}{4} \sum_{i=0,2,4}^4 \binom{4}{i} x^{4-i} (0.5^i + (-0.5)^i) - 0.5 \\
&= (2x + 2) + \frac{1}{4}(x^4 + 3x^2 + 0.125) - 0.5 \\
&= 0.5x^4 + 0.75x^2 + 2x + 1.53125.
\end{aligned}$$

Thus, we have that

$$f_{0.5}(x) = \mathbb{E}_{\tilde{u} \in 0.5\mathbb{B}} [| (x + \tilde{u})^3 + 1 |] = \begin{cases} -x^3 - 0.25x - 1, & \text{if } x \leq -1.5 \\ 0.5x^4 + 0.75x^2 + 2x + 1.53125, & \text{if } x \in [-1.5, -0.5] \\ x^3 + 0.25x + 1, & \text{if } x \geq -0.5. \end{cases}$$

Key observations

Note that $f_{0.5}$ is continuous over its domain, in particular at $x = -0.5$ and $x = -1.5$ with $f_{0.5}(-0.5) = 0.75$ and $f_{0.5}(-1.5) = 2.75$.

Also, it is continuously differentiable with

Preliminary definitions and notation

Let us define \mathbb{S} as the surface of the ball \mathbb{B} , i.e., $\mathbb{S} \triangleq \{v \in \mathbb{R}^n \mid \|v\| = 1\}$.

Also, let $\eta\mathbb{B}$ and $\eta\mathbb{S}$ denote the ball with radius η and its surface, respectively.

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be in the class $C^{k,r}$ if it is k times continuously differentiable and its r th derivative is Lipschitz continuous.

Given a set $X \subseteq \mathbb{R}^n$ and a scalar $\eta > 0$, we let X_η denote the expanded set $X + \eta\mathbb{B}$.

Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and a set $X \subseteq \mathbb{R}^n$, we write $f \in C^{0,0}(X)$ if f is **Lipschitz continuous** on the set X , i.e.,

$$|f(x) - f(\tilde{x})| \leq L_0 \|x - \tilde{x}\|, \quad \text{for all } x, \tilde{x} \in X,$$

and some $L_0 > 0$.

In this case, when $X = \mathbb{R}^n$, we write $f \in C^{0,0}$.

Given a continuously differentiable function and a set $X \subseteq \mathbb{R}^n$, we write $f \in C^{1,1}(X)$ if ∇f is Lipschitz continuous on the set X , i.e., $\|\nabla f(x) - \nabla f(\tilde{x})\| \leq L_1 \|x - \tilde{x}\|$ for all $x, \tilde{x} \in X$ and some $L_1 > 0$.

In this case, when $X = \mathbb{R}^n$, we write $f \in C^{1,1}$.

In-class assignment 2:

Let $X = \{(x_1, x_2) \mid 2 \leq x_1 \leq 4, 1 \leq x_2 \leq 3\}$. In \mathbb{R}^2 , draw the shape of $X_{0,1}$.

In-class assignment 3:

Consider the function $f(x) = |x^3 + 1|$. Is this function Lipschitz continuous?

Properties of spherical smoothing

Lemma 1: Let h_η be defined as $h_\eta(x) \triangleq \mathbb{E}_{u \in \mathbb{B}} [h(x + \eta u)]$. Then the following results hold.

(i) The smoothed function h_η is continuously differentiable over X . In particular, for any $x \in X$, we have that

$$\nabla_x h_\eta(x) = \left(\frac{n}{\eta}\right) \mathbb{E}_{v \in \eta \mathbb{S}} \left[h(x + v) \frac{v}{\|v\|} \right] = \left(\frac{n}{\eta}\right) \mathbb{E}_{v \in \eta \mathbb{S}} \left[(h(x + v) - h(x)) \frac{v}{\|v\|} \right].$$

Suppose $h \in C^{0,0}(X_\eta)$ with parameter L_0 . For any $x, y \in X$, we have that (ii) -- (iv) hold.

(ii) $|h_\eta(x) - h_\eta(y)| \leq L_0 \|x - y\|$.

(iii) $|h_\eta(x) - h(x)| \leq L_0 \eta$.

(iv) $\|\nabla_x h_\eta(x) - \nabla_x h_\eta(y)\| \leq \frac{L_0 n}{\eta} \|x - y\|$.

(v) If h is convex and $h \in C^{0,0}(X_\eta)$ with parameter L_0 , then h_η is convex and satisfies the following for any $x \in X$.

$$h(x) \leq h_\eta(x) \leq h(x) + \eta L_0.$$

(vi) If h is convex and $h \in C^{0,0}(X_\eta)$ with parameter L_0 , then

$$h(y) + \eta L_0 \geq h(x) + \nabla_x h_\eta(x)^T (y - x), \quad \text{for all } y \in X.$$

(vii) If $h \in C^{1,1}(X_\eta)$ with constant L_1 , then $\|\nabla_x h_\eta(x) - \nabla_x h(x)\| \leq \eta L_1 n$.

(viii) Suppose $h \in C^{0,0}(X_\eta)$ with parameter L_0 . Let us define for $v \in \eta \mathbb{S}$

$$g_\eta(x, v) \triangleq \left(\frac{n}{\eta}\right) \frac{(h(x+v) - h(x))v}{\|v\|}.$$

Then, for any $x \in X$, we have that $\mathbb{E}_{v \in \eta \mathbb{S}} [\|g_\eta(x, v)\|^2] \leq L_0^2 n^2$.

Jensen's inequality: Given a random variable Z and a convex function $q(Z)$, then we have $q(\mathbb{E}[Z]) \leq \mathbb{E}[q(Z)]$.

$$Z := h(x + \eta u) - h(y + \eta u)$$

$$q(Z) = |Z|$$

Proof: (For a complete proof, see <https://link.springer.com/article/10.1007/s10107-022-01893-6> (<https://link.springer.com/article/10.1007/s10107-022-01893-6>))

(i) Omitted.

(ii) We have

$$\begin{aligned} |h_\eta(x) - h_\eta(y)| &= |\mathbb{E}_{u \in \mathbb{B}} [h(x + \eta u)] - \mathbb{E}_{u \in \mathbb{B}} [h(y + \eta u)]| \stackrel{\text{Jensen's ineq.}}{\leq} \mathbb{E}_{u \in \mathbb{B}} [|h(x + \eta u) - h(y + \eta u)|] \\ &\stackrel{h \in C^{0,0}(X_\eta)}{\leq} \mathbb{E}_{u \in \mathbb{B}} [L_0 \|x - y\|] = L_0 \|x - y\|. \end{aligned}$$

(iii) Next, we show that $|h_\eta(x) - h(x)|$ can be bounded in terms of η and L_0 .

$$\begin{aligned} |h_\eta(x) - h(x)| &= \left| \int_{\eta \mathbb{B}} (h(x + u) - h(x)) p(u) du \right| \\ &\leq \int_{\eta \mathbb{B}} |(h(x + u) - h(x))| p(u) du \\ &\leq L_0 \int_{\eta \mathbb{B}} \|u\| p(u) du \leq L_0 \eta \int_{\eta \mathbb{B}} p(u) du = L_0 \eta. \end{aligned}$$

(iv) Note that we have $X + \eta \mathbb{S} \subseteq X + \eta \mathbb{B}$. Thus, from the definition of X_η and $h \in C^{0,0}(X_\eta)$, we have $h \in C^{0,0}(X + \eta \mathbb{S})$. As such, we have

$$\begin{aligned} \|\nabla_x h_\eta(x) - \nabla_x h_\eta(y)\| &= \left\| \frac{n}{\eta} \mathbb{E}_{v \in \eta \mathbb{S}} \left[h(x + v) \frac{v}{\|v\|} \right] - \frac{n}{\eta} \mathbb{E}_{v \in \mathbb{S}} \left[h(y + v) \frac{v}{\|v\|} \right] \right\| \\ &\leq \frac{n}{\eta} \mathbb{E}_{v \in \eta \mathbb{S}} \left[\left\| (h(x + v) - h(y + v)) \frac{v}{\|v\|} \right\| \right] \\ &\leq \frac{L_0 n}{\eta} \|x - y\| \mathbb{E}_{v \in \eta \mathbb{S}} \left[\frac{\|v\|}{\|v\|} \right] = \frac{L_0 n}{\eta} \|x - y\|. \end{aligned}$$

(v) Omitted.

(vi) From part (v), function h_η is convex and $h(y) + \eta L_0 \geq h_\eta(y)$ for any $y \in X$. Thus, for all $x, y \in X$ we have

$$h(y) + \eta L_0 \geq h_\eta(y) \geq h_\eta(x) + \nabla h_\eta(x)^T (y - x) \geq h(x) + \nabla h_\eta(x)^T (y - x).$$

(vii) Note that we can show that $\int_{\eta\mathbb{S}} v v^T p_v(v) dv = \frac{\eta^2}{n} \mathbf{I}$. We may then express $\nabla_x h(x)$ as

$$\begin{aligned} \nabla_x h(x) &= \frac{n}{\eta^2} \left(\int_{\eta\mathbb{S}} v v^T p_v(v) dv \right) \nabla_x h(x) = \frac{n}{\eta^2} \left(\int_{\eta\mathbb{S}} v^T \nabla_x h(x) v p_v(v) dv \right) \\ &= \frac{n}{\eta} \left(\int_{\eta\mathbb{S}} v^T \nabla_x h(x) \frac{v}{\|v\|} p_v(v) dv \right) = \frac{n}{\eta} \mathbb{E}_{v \in \eta\mathbb{S}} \left[\left(\nabla_x h(x)^T v \right) \frac{v}{\|v\|} \right], \end{aligned}$$

where the third inequality follows from $\|v\| = \eta$ for $v \in \eta\mathbb{S}$. From this relation, part (i), and by recalling that $\frac{n}{\eta} \mathbb{E}_{v \in \eta\mathbb{S}} \left[h(x) \frac{v}{\|v\|} \right] = 0$, we can write

$$\begin{aligned} \|\nabla_x h_\eta(x) - \nabla_x h(x)\| &= \left\| \frac{n}{\eta} \mathbb{E}_{v \in \eta\mathbb{S}} \left[(h(x+v) - h(x)) \frac{v}{\|v\|} \right] - \frac{n}{\eta} \mathbb{E}_{v \in \eta\mathbb{S}} \left[(\nabla h(x)^T v) \frac{v}{\|v\|} \right] \right\| \\ &\leq \frac{n}{\eta} \mathbb{E}_{v \in \eta\mathbb{S}} \left[\left\| h(x+v) - h(x) - \nabla h(x)^T v \right\| \frac{\|v\|}{\|v\|} \right] \\ &\leq \frac{n}{\eta} \mathbb{E}_{v \in \eta\mathbb{S}} [L_1 \|v\|^2] = n\eta L_1. \end{aligned}$$

(viii) We observe that for any x , $\mathbb{E}_{v \in \eta\mathbb{S}} [\|g_\eta(x, v)\|^2]$ may be bounded as follows.

$$\begin{aligned} \mathbb{E}_{v \in \eta\mathbb{S}} [\|g_\eta(x, v)\|^2] &= \frac{n^2}{\eta^2} \int_{\eta\mathbb{S}} \frac{\|(h(x+v) - h(x))v\|^2}{\|v\|^2} p_v(v) dv \\ &\leq \frac{n^2}{\eta^2} \int_{\eta\mathbb{S}} L_0^2 \|v\|^2 p_v(v) dv \leq n^2 \int_{\eta\mathbb{S}} p_v(v) dv = n^2 L_0^2. \end{aligned}$$

Zeroth-order stochastic gradient descent method: theory and implementation

Problem formulation

We consider solving the following [stochastic optimization](#) problem.

$$\begin{aligned} &\text{minimize} && f(x) \triangleq \mathbb{E}[F(x, \xi)] \\ &\text{subject to:} && \\ &&& x \in X. \end{aligned}$$

(P)

- Here, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is an unknown deterministic function.
- Here, $F : \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}$ is a known stochastic function.
- $\xi : \Omega \rightarrow \mathbb{R}^d$ is a d -dimensional random variable.
- $\mathbb{E}[\bullet]$ denotes the expectation operator with respect to ξ .
- The gradient mapping of F is not available. We only have access to an oracle that returns F at given x and ξ .

To address (P) in the nonconvex constrained case, we consider solving a smoothed approximate problem of the form

$$\begin{aligned} &\text{minimize} && f_\eta(x) \triangleq \mathbb{E}_{u \in \mathbb{B}, \xi} [F(x + \eta u, \xi)] \\ &\text{subject to:} && \\ &&& x \in X. \end{aligned}$$

(P _{η})

Algorithm outline (ZO-SGD for nondifferentiable nonconvex case)

Initialization: Choose a random vector $x_0 \in X$, a stepsize γ , $K \geq 1$, a random value R from $\{0, \dots, K-1\}$,

for $k = 0, \dots, K-1$

Generate random realization of ξ , denoted by ξ_k

Generate random realization of $v \in \eta\mathbb{S}$, denoted by v_k

Evaluate the zeroth-order stochastic gradient as follows.

$$g_\eta(x_k, v_k, \xi_k) = \left(\frac{n}{\eta} \right) (F(x_k + v_k, \xi_k) - F(x_k, \xi_k)) \frac{v_k}{\|v_k\|}.$$

Update the main iterate as follows.

$$x_{k+1} = \mathcal{P}_X \left(x_k - \gamma g_\eta(x_k, v_k, \xi_k) \right).$$

end for

Output: Return x_R

$P_X(\hat{x})$ is the projection of \hat{x} onto the set X .

We let the history of the method of random variables used up to iteration k be denoted by

$$\mathcal{F}_k \triangleq \{x_0, \xi_0, v_0, \xi_1, v_1, \dots, \xi_{k-1}, v_{k-1}\} \quad \text{for all } k \geq 1,$$

and $\mathcal{F}_0 \triangleq \{x_0\}$. We also let $\mathbb{E}[\bullet \mid \mathcal{F}_k]$ denote the conditional expectation with respect to the filtration \mathcal{F}_k .

Assumption 1: Random samples $\{\xi_k\}$ are iid drawn from Ω . Random samples $\{v_k\}$ are iid drawn from $\eta\mathbb{S}$. Also, $\{\xi_k\}$ and $\{v_k\}$ are independent. Moreover, $f(x) = \mathbb{E}[F(x, \xi) \mid x]$ for all $x \in X + \eta_0\mathbb{B}$.

Assumption 1 implies that $F(\bullet, \xi)$ is an unbiased estimator of the true value of the objective function, that is $f(x)$.

Assumption 2: The set $X \subset \mathbb{R}^n$ is closed, convex, and bounded.

Assumption 3: Suppose for any arbitrary $\xi \in \Omega$, we have $F(\bullet, \xi) \in C^{0,0}(X + \eta_0\mathbb{B})$, that is, $F(x, \xi)$ is **Lipschitz continuous** on the set $X + \eta_0\mathbb{B}$, i.e.,

$$|F(x, \xi) - F(\tilde{x}, \xi)| \leq L_0(\xi)\|x - \tilde{x}\|, \quad \text{for all } x, \tilde{x} \in X + \eta_0\mathbb{B},$$

and some $L_0(\xi) > 0$. Also, suppose $L_0 := \sqrt{\mathbb{E}[L_0^2(\xi)]} < \infty$.

Why is $g_\eta(x_k, v_k, \xi_k)$ a suitable stochastic gradient?

Lemma 2 (Properties of the zeroth-order stochastic gradient): Consider the ZO-SGD method. Let Assumptions 1, 2, and 3 hold. Then, we have:

(i) $\mathbb{E}[g_\eta(x_k, v_k, \xi_k) \mid \mathcal{F}_k] = \nabla f_\eta(x_k).$

(ii) $\mathbb{E}[\|g_\eta(x_k, v_k, \xi_k)\|^2 \mid \mathcal{F}_k] \leq L_0^2 n^2.$

Proof: To show part (i), we can write

$$\begin{aligned} \mathbb{E}[g_\eta(x_k, v_k, \xi_k) \mid \mathcal{F}_k] &= \mathbb{E}\left[\left(\frac{n}{\eta}\right) (F(x_k + v_k, \xi_k) - F(x_k, \xi_k)) \frac{v_k}{\|v_k\|} \mid \mathcal{F}_k\right] \\ &\stackrel{\text{Law of total expectation}}{=} \left(\frac{n}{\eta}\right) \mathbb{E}_{v_k} \left[\mathbb{E}_{\xi_k} \left[(F(x_k + v_k, \xi_k) - F(x_k, \xi_k)) \frac{v_k}{\|v_k\|} \mid \mathcal{F}_k \cup \{v_k\} \right] \right] \\ &\stackrel{\text{Assumption 1}}{=} \left(\frac{n}{\eta}\right) \mathbb{E}_{v_k} \left[(f(x_k + v_k) - f(x_k)) \frac{v_k}{\|v_k\|} \mid \mathcal{F}_k \right] \\ &\stackrel{\mathbb{E}[v_k]=0}{=} \left(\frac{n}{\eta}\right) \mathbb{E}_{v_k} \left[f(x_k + v_k) \frac{v_k}{\|v_k\|} \mid \mathcal{F}_k \right] \\ &\stackrel{\text{Lemma 1(i)}}{=} \nabla f_\eta(x_k). \end{aligned}$$

To show part (ii), we can write

$$\begin{aligned} \mathbb{E} \left[\|g_\eta(x_k, v_k, \xi_k)\|^2 \mid \mathcal{F}_k \right] &= \left(\frac{n}{\eta} \right)^2 \mathbb{E} \left[\left\| (F(x_k + v_k, \xi_k) - F(x_k, \xi_k)) \frac{v_k}{\|v_k\|} \right\|^2 \mid \mathcal{F}_k \right] \\ &\stackrel{\text{Law of total expectation}}{=} \left(\frac{n}{\eta} \right)^2 \mathbb{E}_{v_k} \left[\mathbb{E}_{\xi_k} \left[\left\| (F(x_k + v_k, \xi_k) - F(x_k, \xi_k)) \frac{v_k}{\|v_k\|} \right\|^2 \mid \mathcal{F}_k \cup \{v_k\} \right] \right] \\ &= \frac{1}{\|v_k\|^2} \|v_k\|^2 \end{aligned}$$

Definition 1: Let us define the stochastic errors $w_k \triangleq g_\eta(x_k, v_k, \xi_k) - \nabla f_\eta(x_k)$ for all $k \geq 0$.

Note that the main update rule of the algorithm can be rewritten as

$$x_{k+1} = \mathcal{P}_X \left(x_k - \gamma (\nabla f_\eta(x_k) + w_k) \right).$$

The notion of stationarity for nonconvex optimization over a convex set

Lemma 3: Consider the minimization of an L -smooth but possibly nonconvex function over the closed convex set X . Then, x is a stationary point of this problem if and only if

$$\|G_L(x)\| = 0$$

where $G_L(x)$ is called the residual mapping and is defined as

$$G_L(x) \triangleq L \left(x - \mathcal{P}_X \left(x - \frac{1}{L} \nabla_x f(x) \right) \right).$$

Definition 2 (The residual mappings of the smooth problem): Suppose Assumptions 1-3 hold. Given a scalar $\beta > 0$ and a smoothing parameter $\eta > 0$, for any $x \in \mathbb{R}^n$, let the residual mapping $G_{\eta, \beta}$ and its error-afflicted counterpart $\tilde{G}_{\eta, \beta}$ be defined as

$$\begin{aligned} G_{\eta, \beta}(x) &\triangleq \beta \left(x - \mathcal{P}_X \left(x - \frac{1}{\beta} \nabla_x f_\eta(x) \right) \right), \\ \tilde{G}_{\eta, \beta}(x) &\triangleq \beta \left(x - \mathcal{P}_X \left(x - \frac{1}{\beta} (\nabla_x f_\eta(x) + \tilde{e}) \right) \right), \end{aligned}$$

where $\tilde{e} \in \mathbb{R}^n$ is an arbitrary given vector.

Remark 1: Consider the compact representation of the update rule of ZO-SGD. Let us define $\tilde{e} := w_k$.

Note that we have

$$\|x_{k+1} - x_k\|^2 = \gamma^2 \|\tilde{G}_{\eta, \frac{1}{\gamma}}(x_k)\|^2.$$

$$\|x_{k+1} - x_k\|^2 = \|\mathcal{P}_X(x_k - \gamma g_\eta(x_k, v_k, \xi_k)) - x_k\|^2 = \|\mathcal{P}_X(x_k - \gamma (\nabla f_\eta(x_k) + w_k)) - x_k\|^2$$

Lemma 4: Let Assumptions 1-3 hold. Then the following holds for any $\beta > 0$, $\eta > 0$, and $x \in \mathbb{R}^n$.

$$\|G_{\eta, \beta}(x)\|^2 \leq 2\|\tilde{G}_{\eta, \beta}(x)\|^2 + 2\|\tilde{e}\|^2.$$

From Definition 2, we may bound $G_{\eta, \beta}(x)$ as follows.

$$\begin{aligned} \|G_{\eta, \beta}(x)\|^2 &= \left\| \beta \left(x - \mathcal{P}_X \left(x - \frac{1}{\beta} \nabla_x f_\eta(x) \right) \right) \right\|^2 \\ &= \left\| \beta \left(x - \mathcal{P}_X \left(x - \frac{1}{\beta} (\nabla_x f_\eta(x) + \tilde{e}) \right) \right) \right\|^2 \\ &\quad + \left\| \beta \mathcal{P}_X \left(x - \frac{1}{\beta} (\nabla_x f_\eta(x) + \tilde{e}) \right) - \beta \mathcal{P}_X \left(x - \frac{1}{\beta} \nabla_x f_\eta(x) \right) \right\|^2 \\ &\leq 2 \left\| \beta \left(x - \mathcal{P}_X \left(x - \frac{1}{\beta} (\nabla_x f_\eta(x) + \tilde{e}) \right) \right) \right\|^2 \\ &\quad + 2 \left\| \beta \mathcal{P}_X \left(x - \frac{1}{\beta} (\nabla_x f_\eta(x) + \tilde{e}) \right) - \beta \mathcal{P}_X \left(x - \frac{1}{\beta} \nabla_x f_\eta(x) \right) \right\|^2 \\ &\leq 2\|\tilde{G}_{\eta, \beta}(x)\|^2 + 2\|\tilde{e}\|^2, \end{aligned}$$

where the last inequality is a consequence of the non-expansivity of the Euclidean projector.

$$\|u + v\| \leq \|u\| + \|v\|$$

$$\|u + v\|^2 \leq (\|u\| + \|v\|)^2 \leq 2\|u\|^2 + 2\|v\|^2$$

Convergence analysis of ZO-SGD (nonconvex case)

Theorem 1: Let $f^* := \min_{x \in \mathbb{R}^n} f(x)$ and let Assumptions 1,2, and 3 hold. Suppose x_R is generated by the ZO-SGD method. Also, choose γ and η such that $\frac{\gamma n L_0}{\eta} < 0.5$. Then,

$$\mathbb{E} \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq \frac{8(\mathbb{E}[f(x_0)] - f^* + 2L_0\eta)}{\gamma K} + \frac{8 \sum_{k=0}^{K-1} \mathbb{E}[\|w_k\|^2]}{K}.$$

Proof:

Note that from Lemma 1(iv), the function f_η is $\frac{nL_0}{\eta}$ -smooth. Invoking the descent lemma for the smoothed function $f_\eta(x)$ we have

$$f_\eta(x_{k+1}) \leq f_\eta(x_k) + \nabla f_\eta(x_k)^T (x_{k+1} - x_k) + \frac{nL_0}{2\eta} \|x_{k+1} - x_k\|^2.$$

Invoking the projection theorem (see Lecture_notes_20230127 by choosing $\hat{x} := x_k - \gamma(\nabla f_\eta(x_k) + w_k)$), we obtain

$$(x_{k+1} - (x_k - \gamma(\nabla f_\eta(x_k) + w_k)))^T (x - x_{k+1}) \geq 0, \quad \forall x \in X.$$

Let us choose $x := x_k$. Then we have

$$(x_{k+1} - x_k + \gamma(\nabla f_\eta(x_k) + w_k))^T (x_k - x_{k+1}) \geq 0.$$

Thus we have

$$-\|x_k - x_{k+1}\|^2 + \gamma w_k^T (x_k - x_{k+1}) \geq \nabla f_\eta(x_k)^T (x_{k+1} - x_k).$$

Rearranging the terms we obtain

$$\nabla f_\eta(x_k)^T (x_{k+1} - x_k) \leq -\frac{1}{\gamma} \|x_k - x_{k+1}\|^2 + w_k^T (x_k - x_{k+1}).$$

From the preceding relation and the first inequality we obtain

$$f_\eta(x_{k+1}) \leq f_\eta(x_k) - \frac{1}{\gamma} \|x_{k+1} - x_k\|^2 + w_k^T (x_k - x_{k+1}) + \frac{nL_0}{2\eta} \|x_{k+1} - x_k\|^2.$$

We can write $w_k^T (x_k - x_{k+1}) \leq \frac{\gamma}{2} \|w_k\|^2 + \frac{1}{2\gamma} \|x_{k+1} - x_k\|^2$. This implies that

$$\begin{aligned} f_\eta(x_{k+1}) &\leq f_\eta(x_k) - \frac{1}{2\gamma} \|x_{k+1} - x_k\|^2 + \frac{\gamma}{2} \|w_k\|^2 + \frac{nL_0}{2\eta} \|x_{k+1} - x_k\|^2 \\ &= f_\eta(x_k) - \frac{1}{2\gamma} \left(1 - \frac{\gamma n L_0}{\eta}\right) \|x_{k+1} - x_k\|^2 + \frac{\gamma}{2} \|w_k\|^2. \end{aligned}$$

Invoking Remark 1, we obtain

$$f_\eta(x_{k+1}) \leq f_\eta(x_k) - \frac{1}{2\gamma} \left(1 - \frac{\gamma n L_0}{\eta}\right) \gamma^2 \|\tilde{G}_{\eta, \frac{1}{\gamma}}(x_k)\|^2 + \frac{\gamma}{2} \|w_k\|^2.$$

Suppose $\frac{\gamma n L_0}{\eta} < 1$. Invoking Lemma 4, we get

$$\begin{aligned} f_\eta(x_{k+1}) &\leq f_\eta(x_k) - \frac{1}{2\gamma} \left(1 - \frac{\gamma n L_0}{\eta}\right) \gamma^2 (0.5) \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 + \frac{1}{2\gamma} \left(1 - \frac{\gamma n L_0}{\eta}\right) \gamma^2 \|w_k\|^2 + \frac{\gamma}{2} \|w_k\|^2 \\ &\leq f_\eta(x_k) - \frac{\gamma}{4} \left(1 - \frac{\gamma n L_0}{\eta}\right) \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 + \gamma \|w_k\|^2. \end{aligned}$$

Suppose $\frac{\gamma n L_0}{\eta} < 0.5$. Then, $\left(1 - \frac{\gamma n L_0}{\eta}\right) > 0.5$, implying that

$$f_\eta(x_{k+1}) \leq f_\eta(x_k) - \frac{\gamma}{8} \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 + \gamma \|w_k\|^2.$$

Thus we have

$$\frac{\gamma}{8} \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 \leq f_\eta(x_k) - f_\eta(x_{k+1}) + \gamma \|w_k\|^2.$$

Taking sum on both sides for $k = 0, \dots, K-1$ we obtain

$$\frac{\gamma}{8} \sum_{k=0}^{K-1} \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 \leq f_\eta(x_0) - f_\eta(x_K) + \gamma \sum_{k=0}^{K-1} \|w_k\|^2.$$

Also, note that from the definition of x_R we can write

$$\sum_{k=0}^{K-1} \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 = (K - k_0) \mathbb{E}_R \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right].$$

We obtain

$$\frac{\gamma K}{8} \mathbb{E}_R \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq f_\eta(x_0) - f_\eta(x_K) + \gamma \sum_{k=0}^{K-1} \|w_k\|^2.$$

Invoking Lemma 1(iii) we have

$$|f_\eta(x) - f(x)| \leq L_0 \eta \quad \Rightarrow \quad f(x) - L_0 \eta \leq f_\eta(x) \leq f(x) + L_0 \eta.$$

We obtain

$$\frac{\gamma K}{8} \mathbb{E}_R \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq f(x_0) - f(x_K) + 2L_0 \eta + \gamma \sum_{k=0}^{K-1} \|w_k\|^2.$$

We have

$$\mathbb{E}_R \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq \frac{8(f(x_0) - f^* + 2L_0 \eta)}{\gamma K} + \frac{8 \sum_{k=0}^{K-1} \|w_k\|^2}{K}.$$

Taking expectation on both sides (with respect to \mathcal{F}_K), we have

$$\mathbb{E} \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq \frac{8(\mathbb{E}[f(x_0)] - f^* + 2L_0 \eta)}{\gamma K} + \frac{8 \sum_{k=0}^{K-1} \mathbb{E}[\|w_k\|^2]}{K}.$$

Analyzing the term $\mathbb{E}[\|w_k\|^2]$

Lemma 5: Consider Definition 1. The following hold for all $k \geq 0$.

(i) $\mathbb{E}[w_k \mid \mathcal{F}_k] = 0$.

(ii) $\mathbb{E}[\|w_k\|^2] \leq L_0^2 n^2$.

Proof:

Part (i) is implied by Lemma 2(i).

Part (ii) is shown as follows.

$$\begin{aligned} \mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] &= \mathbb{E}[\|g_\eta(x_k, v_k, \xi_k) - \nabla f_\eta(x_k)\|^2 \mid \mathcal{F}_k] = \mathbb{E}[\|g_\eta(x_k, v_k, \xi_k)\|^2 \mid \mathcal{F}_k] + \|\nabla f_\eta(x_k)\|^2 - 2\mathbb{E}[g_\eta(x_k, v_k, \xi_k)^T \nabla f_\eta(x_k) \mid \mathcal{F}_k] \\ &\stackrel{\text{Lemma 2(i)}}{=} \mathbb{E}[\|g_\eta(x_k, v_k, \xi_k)\|^2 \mid \mathcal{F}_k] - \|\nabla f_\eta(x_k)\|^2 \\ &\leq \mathbb{E}[\|g_\eta(x_k, v_k, \xi_k)\|^2 \mid \mathcal{F}_k] \\ &\stackrel{\text{Lemma 2(ii)}}{\leq} L_0^2 n^2. \end{aligned}$$

Applying the law of total expectation, we have $\mathbb{E}[\|w_k\|^2] = \mathbb{E}_{\xi_k, v_k}[\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k]] \leq L_0^2 n^2$.

Remark: The fact that $\mathbb{E}[\|w_k\|^2]$ is bounded by a persistent constant term makes it hard to establish the convergence of the ZO-SGD method. Indeed, we need to employ "variance-reduciton" to be able to decrease the value of this variance. This will be discussed at length in a separate module in this course.

Stochastic Optimization (Module 6: Randomized block methods)

Module's main topics/objectives:

- What is the randomized block coordinate technique?
- How to design a randomized-block SGD method?
- Theoretically, how fast does Randomized-Block SGD converge?
- Numerically, how does it perform on MNIST dataset?

Randomized block-coordinate stochastic gradient descent method

Problem formulation

We consider solving the following [stochastic optimization](#) problem.

$$\begin{array}{ll} \text{minimize} & f(x) \triangleq \mathbb{E}[F(x, \xi)] \\ \text{subject to:} & \\ & x \in X. \end{array}$$

(P)

- Here, $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is an unknown deterministic function.
- Here, $F: \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}$ is a known stochastic function.
- $\xi: \Omega \rightarrow \mathbb{R}^d$ is a d -dimensional random variable.
- $\mathbb{E}[\bullet]$ denotes the expectation operator with respect to ξ .

Assumption 1:

- $\mathbb{E}_\xi[\nabla F(x, \xi) \mid x] = \nabla f(x)$ for all $x \in X$.
- $\mathbb{E}[\|\nabla F(x, \xi) - \nabla f(x)\|^2 \mid x] \leq \sigma^2$ for all $x \in X$ for some $\sigma > 0$.

Assumption 2:

- The set X is given as $X \triangleq \prod_{i=1}^N X_i$ where $X_i \subseteq \mathbb{R}^{n_i}$ and $\sum_{i=1}^N n_i = n$.
- The set X_i is nonempty, compact, and convex for all $i = 1, \dots, N$.

Assumption 3: At iteration $k \geq 0$, a random variable i_k is generated from an independent and identically distributed discrete probability distribution such that $\text{Prob}(i_k = i) = p_i$ where $p_i > 0$ for all $i \in [N]$ and $\sum_{i=1}^N p_i = 1$.

For example, if we have 3 blocks, one can consider a discrete probability distribution as follows.

$$p_1 = 0.5, p_2 = 0.1, p_3 = 0.4$$

$$X_1 = \{x_1 \in \mathbb{R} \mid 1 \leq x_1 \leq 2\}$$

$$X_2 = \{x_2 \in \mathbb{R} \mid 2 \leq x_2 \leq 5\}$$

$$X_1 \times X_2 = \left\{ (x_1, x_2) \mid \begin{bmatrix} 1 \\ 2 \end{bmatrix} \leq \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 2 \\ 5 \end{bmatrix} \right\}$$

$$n_1 = 1$$

$$n_2 = 1$$

$$N = 2$$

$$n = 2$$

$$\mathcal{P}_{X_1}(3) = 2$$

$$\mathcal{P}_{X_1}(1.5) = 1.5$$

$$\mathcal{P}_{X_1}(-2) = 1$$

$$\mathcal{P}_{X_1}(2) = 2$$

$$\mathcal{P}_X([5, 3]) = [2, 3]$$

```
In [20]: 1 import numpy as np
          2
          3 blocks = [1, 2, 3]
          4 probabilities = [0.3, 0.3, .4]
          5
          6 ik=np.random.choice(blocks, 1, p=probabilities)[0]
          7 print(ik)
```

3

Block structure of vector $x \in \mathbb{R}^n$ is as follows.

$$x = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(N)} \end{bmatrix}.$$

Note that we can consider a block structure for ∇f aligned with the block structure of the set X as follows.

$$\nabla f(x) = \begin{bmatrix} \nabla_{x^{(1)}} f(x) \\ \nabla_{x^{(2)}} f(x) \\ \vdots \\ \nabla_{x^{(N)}} f(x) \end{bmatrix},$$

where $\nabla_{x^{(i)}} f(x) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_i}$ denotes the i -th block-coordinate of ∇f .

Similarly, we consider

$$\nabla F(x, \xi) = \begin{bmatrix} \nabla_{x^{(1)}} F(x, \xi) \\ \nabla_{x^{(2)}} F(x, \xi) \\ \vdots \\ \nabla_{x^{(N)}} F(x, \xi) \end{bmatrix},$$

Note that under Assumption 2, we can rewrite problem (P) as

$$\begin{array}{ll} \text{minimize} & f(x) \triangleq \mathbb{E} [F(x^{(1)}, \dots, x^{(N)}), \xi] \\ \text{subject to:} & \\ & x^{(1)} \in X_1 \\ & x^{(2)} \in X_2 \\ & \vdots \\ & x^{(N)} \in X_N \end{array}$$

Algorithm outline (RB-SGD)

Initialization: Choose a random vector $x_0 \in X$, a stepsize sequence $\{\gamma_k\}$, $K \geq 1$ for $k = 0, \dots, K - 1$

Generate a random realization of ξ , denoted by ξ_k

Generate a random realization of block index, denoted by i_k , drawn from $\{1, \dots, N\}$ where $\mathbb{P}(i_k = i) = p_i$ for all $i \in \{1, \dots, N\}$

Update the iterate x_k as follows. For all i ,

$$x_{k+1}^{(i)} := \begin{cases} \mathcal{P}_{X_i} \left(x_k^{(i)} - \gamma_k \nabla_{x^{(i)}} F(x_k, \xi_k) \right), & \text{if } i_k = i \\ x_k^{(i)}, & \text{if } i_k \neq i \end{cases}$$

end for

Output: Return x_K

Lemma 1: Let Assumption 2 hold. Then, for any $x \in X$, we have

$$\mathcal{P}_X(x) = \begin{bmatrix} \mathcal{P}_{X_1}(x^{(1)}) \\ \mathcal{P}_{X_2}(x^{(2)}) \\ \vdots \\ \mathcal{P}_{X_N}(x^{(N)}) \end{bmatrix}.$$

In-class assignment 1: Prove this lemma.

Compact representation of RB-SGD

The update rule of the algorithm implies that

$$x_{k+1} := \begin{bmatrix} x_k^{(1)} \\ x_k^{(2)} \\ \vdots \\ \mathcal{P}_{X_{(i_k)}}(x_k^{(i_k)} - \gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k)) \\ \vdots \\ x_k^{(N)} \end{bmatrix} = \begin{bmatrix} \mathcal{P}_{X_1}(x_k^{(1)}) \\ \mathcal{P}_{X_2}(x_k^{(2)}) \\ \vdots \\ \mathcal{P}_{X_{(i_k)}}(x_k^{(i_k)} - \gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k)) \\ \vdots \\ \mathcal{P}_{X_N}(x_k^{(N)}) \end{bmatrix} = \begin{bmatrix} \mathcal{P}_{X_1}(x_k^{(1)} - \gamma_k 0_{n_1 \times 1}) \\ \mathcal{P}_{X_2}(x_k^{(2)} - \gamma_k 0_{n_2 \times 1}) \\ \vdots \\ \mathcal{P}_{X_{(i_k)}}(x_k^{(i_k)} - \gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k)) \\ \vdots \\ \mathcal{P}_{X_N}(x_k^{(N)} - \gamma_k 0_{n_N \times 1}) \end{bmatrix}.$$

Invoking Lemma 1, we obtain

$$x_{k+1} = \mathcal{P}_X \left(\begin{bmatrix} x_k^{(1)} - \gamma_k 0_{n_1 \times 1} \\ x_k^{(2)} - \gamma_k 0_{n_2 \times 1} \\ \vdots \\ x_k^{(i_k)} - \gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k) \\ \vdots \\ x_k^{(N)} - \gamma_k 0_{n_N \times 1} \end{bmatrix} \right) = \mathcal{P}_X \left(\begin{bmatrix} x_k^{(1)} \\ x_k^{(2)} \\ \vdots \\ x_k^{(i_k)} \\ \vdots \\ x_k^{(N)} \end{bmatrix} - \gamma_k \begin{bmatrix} 0_{n_1 \times 1} \\ 0_{n_2 \times 1} \\ \vdots \\ \nabla_{x^{(i_k)}} F(x_k, \xi_k) \\ \vdots \\ 0_{n_N \times 1} \end{bmatrix} \right).$$

We have

$$x_{k+1} = \mathcal{P}_X \left(x_k - \gamma_k \begin{bmatrix} \mathbf{0}_{n_1 \times n_{i_k}} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \\ \mathbf{0}_{n_2 \times n_{i_k}} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \\ \vdots \\ \mathbf{I}_{n_{i_k} \times n_{i_k}} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \\ \vdots \\ \mathbf{0}_{n_N \times n_{i_k}} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \end{bmatrix} \right) = \mathcal{P}_X \left(x_k - \gamma_k \begin{bmatrix} \mathbf{0}_{n_1 \times n_{i_k}} \\ \mathbf{0}_{n_2 \times n_{i_k}} \\ \vdots \\ \mathbf{I}_{n_{i_k} \times n_{i_k}} \\ \vdots \\ \mathbf{0}_{n_N \times n_{i_k}} \end{bmatrix} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \right).$$

Definition (Operator \mathbf{U}_i):

Let $\mathbf{U}_i \in \mathbb{R}^{n \times n_i}$ for $i \in [N]$ be the collection of matrices such that $\mathbf{I}_n = [\mathbf{U}_1, \dots, \mathbf{U}_N] \in \mathbb{R}^{n \times n}$.

Invoking this definition we can write

$$x_{k+1} = \mathcal{P}_X \left(x_k - \gamma_k \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \right).$$

In-class assignment 2:

Let $N = 2$, $n_1 = 2$, and $n_2 = 5$. Suppose $\nabla F(x_k, \xi_k) = \begin{bmatrix} 2 \\ 3 \\ 6 \\ -2 \\ -1 \\ 3 \\ 91 \end{bmatrix}$.

Write the following terms:

(i) $\nabla_{x^{(1)}} F(x_k, \xi_k)$

(ii) $\nabla_{x^{(2)}} F(x_k, \xi_k)$

(iii) \mathbf{I}_7

(iv) \mathbf{U}_1

(v) \mathbf{U}_2

(vi) $\mathbf{U}_1 \nabla_{x^{(1)}} F(x_k, \xi_k)$

(vii) $\mathbf{U}_2 \nabla_{x^{(2)}} F(x_k, \xi_k)$

Answer: Note that $n = \sum_{i=1}^N n_i = 7$.

(i) $\nabla_{x^{(1)}} F(x_k, \xi_k) = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

(ii) $\nabla_{x^{(2)}} F(x_k, \xi_k) = \begin{bmatrix} 6 \\ -2 \\ -1 \\ 3 \\ 91 \end{bmatrix}$

(iii) $\mathbf{I}_7 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$

(iv) $\mathbf{U}_1 \in \mathbb{R}^{7 \times 2}$

$\mathbf{U}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$

(v) $\mathbf{U}_2 \in \mathbb{R}^{7 \times 5}$

$\mathbf{U}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$

$$\mathbf{U}_1 \nabla_{x^{(1)}} F(x_k, \xi_k) = \begin{bmatrix} 2 \\ 3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 6 \end{bmatrix}$$

In-class assignment 3: Use the definitions Δ_k and w_k to write $\mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k)$ in terms of $\nabla f(x_k)$ added by some noise.

Randomized block errors: Let us define the randomized block errors for $k \geq 0$ as

$$\Delta_k \triangleq p_{i_k}^{-1} \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) - \nabla F(x_k, \xi_k).$$

Stochastic errors: Let us define the stochastic errors for $k \geq 0$ as

$$w_k \triangleq \nabla F(x_k, \xi_k) - \nabla f(x_k).$$

$$\Delta_k + w_k = p_{i_k}^{-1} \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) - \nabla F(x_k, \xi_k) + \nabla F(x_k, \xi_k) - \nabla f(x_k) = p_{i_k}^{-1} \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) - \nabla f(x_k)$$

$$p_{i_k}^{-1} \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) = \nabla f(x_k) + \Delta_k + w_k$$

Then we have

$$\mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) = p_{i_k} (\nabla f(x_k) + \Delta_k + w_k).$$

$$\mathbb{E}[Z] = \int_{\mathcal{Z}} z p(z) dz$$

$$\mathbb{E}[i_k] = \sum_{i=1}^N i p_i$$

$$\mathbb{E}[h(i_k)] = \sum_{i=1}^N h(i) p_i$$

Notation: We let the history of the method of random variables used up to iteration k be denoted by:

$$\mathcal{F}_k \triangleq \{x_0, \xi_0, i_0, \xi_1, i_1, \dots, \xi_{k-1}, i_{k-1}\} \quad \text{for all } k \geq 1,$$

and $\mathcal{F}_0 \triangleq \{x_0\}$. We also let $\mathbb{E}[\bullet \mid \mathcal{F}_k]$ denote the conditional expectation with respect to the filtration \mathcal{F}_k .

Note that knowing \mathcal{F}_k , we would know x_k . Here we say x_k is \mathcal{F}_k -measurable.

In-class assignment 2:

Find the following terms.

$$\mathbb{E}[w_k \mid \mathcal{F}_k] = \mathbb{E}_{\xi_k, i_k} [w_k \mid \mathcal{F}_k] = \mathbb{E}_{i_k} [\mathbb{E}_{\xi_k} [w_k \mid \mathcal{F}_k \cup \{i_k\}]] = \mathbb{E}_{i_k} [0] = 0$$

$$\mathbb{E}[w_k] = \mathbb{E}_{\mathcal{F}_k} [\mathbb{E}_{\xi_k, i_k} [w_k \mid \mathcal{F}_k]] = 0$$

$$\begin{aligned} \mathbb{E}[\Delta_k \mid \mathcal{F}_k \cup \{\xi_k\}] &= \mathbb{E} [p_{i_k}^{-1} \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) - \nabla F(x_k, \xi_k) \mid \mathcal{F}_k \cup \{\xi_k\}] = \mathbb{E} [p_{i_k}^{-1} \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \mid \mathcal{F}_k \cup \{\xi_k\}] - \nabla F(x_k, \xi_k) = \\ &= \sum_{i=1}^N p_i^{-1} \mathbf{U}_i \nabla_{x^{(i)}} F(x_k, \xi_k) p_i - \nabla F(x_k, \xi_k) \end{aligned}$$

$$= \left(\sum_{i=1}^N \mathbf{U}_i \nabla_{x^{(i)}} F(x_k, \xi_k) \right) - \nabla F(x_k, \xi_k) = \nabla F(x_k, \xi_k) - \nabla F(x_k, \xi_k) = 0$$

$$\mathbb{E}[\Delta_k \mid \mathcal{F}_k]$$

$$\mathbb{E}[\Delta_k]$$

```
In [28]: 1 import numpy as np
          2 n = 10**100
          3 a = np.random.randint(1,size=n)

-----
ValueError                                Traceback (most recent call last)
/var/folders/vx/8xj88f1j3l9g2l166w22zxx40000gn/T/ipykernel_59937/3092069270.py in <module>
      1 import numpy as np
      2 n = 10**100
----> 3 a = np.random.randint(1,size=n)

mtrand.pyx in numpy.random.mtrand.RandomState.randint()

_bounded_integers.pyx in numpy.random._bounded_integers._rand_int64()

ValueError: Maximum allowed dimension exceeded
```

In-class assignment 3: Use the definitions Δ_k and w_k to write $\mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k)$ in terms of $\nabla f(x_k)$ added by some noise.

Lemma 2:

$$\mathbb{E} \left[\left\| p_{i_k}^{-1} \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \right\|^2 \right] \leq p_{\min}^{-1} (\sigma^2 + C^2).$$

Proof: Note that knowing \mathcal{F}_k , x_k would be known but $\nabla F(x_k, \xi_k)$ would not be known (why?). Also, $F_{x^{(i_k)}} F(x_k, \xi_k)$ would not be known (why?). But knowing $\mathcal{F}_k \cup \{\xi_k\}$, both x_k and $F(x_k, \xi_k)$ would be known. Taking these into account, we have

$$\begin{aligned} \mathbb{E} \left[\left\| p_{i_k}^{-1} \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \right\|^2 \mid \mathcal{F}_k \cup \{\xi_k\} \right] &= \sum_{i=1}^N p_i \left\| p_i^{-1} \mathbf{U}_i \nabla_{x^{(i)}} F(x_k, \xi_k) \right\|^2 \\ &= \sum_{i=1}^N p_i (p_i^{-2} \left\| \mathbf{U}_i \nabla_{x^{(i)}} F(x_k, \xi_k) \right\|^2) \\ &\leq p_{\min}^{-1} \sum_{i=1}^N \left\| \mathbf{U}_i \nabla_{x^{(i)}} F(x_k, \xi_k) \right\|^2 \\ &= p_{\min}^{-1} \sum_{i=1}^N \left\| \nabla_{x^{(i)}} F(x_k, \xi_k) \right\|^2 \\ &= p_{\min}^{-1} \left\| \nabla F(x_k, \xi_k) \right\|^2. \end{aligned}$$

$$\begin{aligned} \mathbb{E}_{\xi_k} \left[\mathbb{E} \left[\left\| p_{i_k}^{-1} \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \right\|^2 \mid \mathcal{F}_k \cup \{\xi_k\} \right] \right] &= \mathbb{E} \left[\left\| p_{i_k}^{-1} \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \right\|^2 \mid \mathcal{F}_k \right] \leq p_{\min}^{-1} \mathbb{E} \left[\left\| \nabla F(x_k, \xi_k) \right\|^2 \mid \mathcal{F}_k \right] \\ &= p_{\min}^{-1} \mathbb{E} \left[\left\| \nabla F(x_k, \xi_k) - \nabla f(x_k) + \nabla f(x_k) \right\|^2 \mid \mathcal{F}_k \right] \\ &= p_{\min}^{-1} \left(\mathbb{E} \left[\left\| \nabla F(x_k, \xi_k) - \nabla f(x_k) \right\|^2 \mid \mathcal{F}_k \right] + \left\| \nabla f(x_k) \right\|^2 + 2 \nabla f(x_k)^T \mathbb{E} \left[\nabla F(x_k, \xi_k) - \nabla f(x_k) \mid \mathcal{F}_k \right] \right) \\ &= p_{\min}^{-1} \left(\mathbb{E} \left[\left\| \nabla F(x_k, \xi_k) - \nabla f(x_k) \right\|^2 \mid \mathcal{F}_k \right] + \left\| \nabla f(x_k) \right\|^2 \right) \end{aligned}$$

where such $C > 0$ exists because of continuity of F over the compact set X .

Compact representation of the RB-SGD method

Given $i_k \in [N]$, for all $i \in [N]$ we have

$$x_{k+1}^{(i)} = \begin{cases} \mathcal{P}_{X_i} \left(x_k^{(i)} - \gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k) \right) & \text{if } i = i_k \\ x_k^{(i)} & \text{otherwise.} \end{cases}$$

Equivalently, we obtained

$$x_{k+1} = \mathcal{P}_X \left(x_k - \gamma_k \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k) \right).$$

Or equivalently,

$$x_{k+1} = \mathcal{P}_X \left(x_k - p_{i_k} \gamma_k (\nabla F(x_k, \xi_k) + \Delta_k) \right).$$

Or equivalently,

$$x_{k+1} = \mathcal{P}_X \left(x_k - p_{i_k} \gamma_k (\nabla f(x_k) + w_k + \Delta_k) \right).$$

Output of the method in the convex case

Definition: Let us define the weighted averaging sequence as follows.

$$\bar{x}_K \triangleq \sum_{k=0}^K \alpha_{k,K} x_k, \quad \text{for } K \geq 0,$$

where $\alpha_{k,K} \triangleq \frac{\gamma_k^r}{\sum_{t=0}^K \gamma_t^r}$ for $k = 0, \dots, K$, and $r \in [0, 1)$.

Distance function: For any $x, y \in \mathbb{R}^n$, function $D : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is defined as

$$D(x, y) \triangleq \sum_{i=1}^N p_i^{-1} \|x^{(i)} - y^{(i)}\|^2.$$

Remark 1: Let $p_{\min} \triangleq \min_{i \in [N]} p_i$ and $p_{\max} \triangleq \max_{i \in [N]} p_i$. Note that we have

$$p_{\max}^{-1} \|x - y\|^2 \leq D(x, y) \leq p_{\min}^{-1} \|x - y\|^2,$$

and

$$p_{\min} D(x, y) \leq \|x - y\|^2 \leq p_{\max} D(x, y).$$

Theorem 1: Consider problem (SVI) and Algorithm 1. Let Assumption 1, 2, and 3 hold. Let us define the weighted averaging sequence as follows.

$$\bar{x}_K \triangleq \sum_{k=0}^K \alpha_{k,K} x_k, \quad \text{for } K \geq 0,$$

where $\alpha_{k,K} \triangleq \frac{\gamma_k^r}{\sum_{i=0}^K \gamma_i^r}$ for $k = 0, \dots, K$, and $r \in [0, 1)$.

(a) Assume that $\{\gamma_k\}$ is a nonnegative and non-increasing sequence. Then, we have for all $K \geq 0$

$$\mathbb{E}[f(\bar{x}_K)] - f^* \leq \frac{0.5 p_{\min}^{-1} M \gamma_K^{r-1} + 0.5 p_{\min}^{-1} (C^2 + \sigma^2) \sum_{k=0}^K \gamma_k^{r+1}}{\sum_{k=0}^K \gamma_k^r}.$$

(b) Let $r := 0$ and the step-size sequence be given by $\gamma_k \triangleq \frac{\gamma_0}{\sqrt{k+1}}$ for all $k \geq 0$. We have for all $K \geq 3$

$$\mathbb{E}[f(\bar{x}_K)] - f^* \leq \left(0.5 \frac{p_{\min}^{-1} M}{\gamma_0} + \gamma_0 p_{\min}^{-1} (C^2 + \sigma^2) \right) \frac{1}{\sqrt{K+1}}.$$

Proof:

(a)

From the update rule of the algorithm and the definition of the distance function we have

$$\begin{aligned} D(x_{k+1}, y) &= \sum_{i=1}^N p_i^{-1} \|x_{k+1}^{(i)} - y^{(i)}\|^2 \\ &= p_{i_k}^{-1} \|x_{k+1}^{(i_k)} - y^{(i_k)}\|^2 + \sum_{i=1, i \neq i_k}^N p_i^{-1} \|x_{k+1}^{(i)} - y^{(i)}\|^2 \\ &= p_{i_k}^{-1} \left\| \mathcal{P}_{X_{i_k}} \left(x_k^{(i_k)} - \gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k) \right) - \mathcal{P}_{X_{i_k}} \left(y^{(i_k)} \right) \right\|^2 + \sum_{i=1, i \neq i_k}^N p_i^{-1} \|x_{k+1}^{(i)} - y^{(i)}\|^2 \\ &\leq p_{i_k}^{-1} \left\| x_k^{(i_k)} - \gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k) - y^{(i_k)} \right\|^2 + \sum_{i=1, i \neq i_k}^N p_i^{-1} \|x_{k+1}^{(i)} - y^{(i)}\|^2, \end{aligned}$$

where we used $y^{(i_k)} \in X_{i_k}$ (see Lemma 1) and the nonexpansiveness property of the projection operator. We obtain

$$\begin{aligned} D(x_{k+1}, y) &\leq p_{i_k}^{-1} \left(\|x_k^{(i_k)} - y^{(i_k)}\|^2 + \gamma_k^2 \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 - 2\gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k)^T (x_k^{(i_k)} - y^{(i_k)}) \right) + \sum_{i=1, i \neq i_k}^N p_i^{-1} \|x_{k+1}^{(i)} - y^{(i)}\|^2 \\ &= p_{i_k}^{-1} \left(\|x_k^{(i_k)} - y^{(i_k)}\|^2 + \gamma_k^2 \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 - 2\gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k)^T (x_k^{(i_k)} - y^{(i_k)}) \right) + \sum_{i=1, i \neq i_k}^N p_i^{-1} \|x_k^{(i)} - y^{(i)}\|^2 \\ &= \sum_{i=1}^N p_i^{-1} \|x_k^{(i)} - y^{(i)}\|^2 + p_{i_k}^{-1} \left(\gamma_k^2 \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 - 2\gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k)^T (x_k^{(i_k)} - y^{(i_k)}) \right) \\ &= D(x_k, y) + p_{i_k}^{-1} \gamma_k^2 \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 - 2p_{i_k}^{-1} \gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k)^T (x_k^{(i_k)} - y^{(i_k)}). \end{aligned}$$

Next, we consider the term $p_{i_k}^{-1} \gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k)^T (x_k^{(i_k)} - y^{(i_k)})$. We have

$$\begin{aligned} p_{i_k}^{-1} \gamma_k \nabla_{x^{(i_k)}} F(x_k, \xi_k)^T (x_k^{(i_k)} - y^{(i_k)}) &= p_{i_k}^{-1} \gamma_k (x_k^{(i_k)} - y^{(i_k)})^T \nabla_{x^{(i_k)}} F(x_k, \xi_k) \\ &= p_{i_k}^{-1} \gamma_k (x_k - y)^T (\mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k)) \\ &= \gamma_k (x_k - y)^T (p_{i_k}^{-1} \mathbf{U}_{i_k} \nabla_{x^{(i_k)}} F(x_k, \xi_k)) \\ &= \gamma_k (x_k - y)^T (\nabla F(x_k, \xi_k) + \Delta_k) \\ &= \gamma_k (x_k - y)^T (\nabla f(x_k) + w_k + \Delta_k). \end{aligned}$$

We obtain

$$D(x_{k+1}, y) \leq D(x_k, y) + p_{i_k}^{-1} \gamma_k^2 \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 - 2\gamma_k(x_k - y)^T \nabla f(x_k) + 2\gamma_k(y - x_k)^T (w_k + \Delta_k).$$

We have

$$\gamma_k(x_k - y)^T \nabla f(x_k) \leq 0.5D(x_k, y) - 0.5D(x_{k+1}, y) + 0.5p_{i_k}^{-1} \gamma_k^2 \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 + \gamma_k(y - x_k)^T (w_k + \Delta_k).$$

Note that the from the convexity of f we can write

$$f(y) - f(x_k) \geq (y - x_k)^T \nabla f(x_k) \Rightarrow f(x_k) - f(y) \geq (x_k - y)^T \nabla f(x_k).$$

From the two preceding relations we obtain

$$\gamma_k(f(x_k) - f(y)) \leq 0.5D(x_k, y) - 0.5D(x_{k+1}, y) + 0.5p_{i_k}^{-1} \gamma_k^2 \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 + \gamma_k(y - x_k)^T (w_k + \Delta_k).$$

Multiplying both sides by γ_k^{r-1} , we have

$$\gamma_k^r(f(x_k) - f(y)) \leq 0.5\gamma_k^{r-1}(D(x_k, y) - D(x_{k+1}, y)) + 0.5p_{i_k}^{-1} \gamma_k^{1+r} \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 + \gamma_k^r(y - x_k)^T (w_k + \Delta_k). \quad (1)$$

Adding and subtracting the term $\frac{\gamma_{k-1}^{r-1}}{2} D(x_k, y)$, we have

$$\begin{aligned} \gamma_k^r(f(x_k) - f(y)) &\leq \frac{\gamma_{k-1}^{r-1}}{2}(D(x_k, y) - D(x_{k+1}, y)) + \frac{\gamma_k^{r-1}}{2} D(x_k, y) - \frac{\gamma_{k-1}^{r-1}}{2} D(x_k, y) \\ &\quad + 0.5p_{i_k}^{-1} \gamma_k^{1+r} \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 + \gamma_k^r(y - x_k)^T (w_k + \Delta_k). \end{aligned}$$

From compactness of X , there exists $M > 0$ such that $\sup_{x \in X} \|x - y\|^2 \leq M$ for all $x, y \in X$. Thus, from Remark 1 we have

$$D(x_k, y) \leq p_{\min}^{-1} \|x_k - y\|^2 \leq p_{\min}^{-1} M.$$

Note that since $r < 1$ and γ_k is nonincreasing, we have $\frac{\gamma_k^{r-1}}{2} - \frac{\gamma_{k-1}^{r-1}}{2} > 0$. Thus we obtain

$$\frac{\gamma_k^{r-1}}{2} D(x_k, y) - \frac{\gamma_{k-1}^{r-1}}{2} D(x_k, y) \leq 0.5(\gamma_k^{r-1} - \gamma_{k-1}^{r-1}) p_{\min}^{-1} M.$$

Thus we obtain

$$\begin{aligned} \gamma_k^r(f(x_k) - f(y)) &\leq \frac{\gamma_{k-1}^{r-1}}{2}(D(x_k, y) - D(x_{k+1}, y)) + 0.5(\gamma_k^{r-1} - \gamma_{k-1}^{r-1}) p_{\min}^{-1} M \\ &\quad + 0.5p_{i_k}^{-1} \gamma_k^{1+r} \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 + \gamma_k^r(y - x_k)^T (w_k + \Delta_k). \end{aligned}$$

Summing from $k = 1, \dots, K$ we obtain

$$\begin{aligned} \sum_{k=1}^K \gamma_k^r(f(x_k) - f(y)) &\leq \sum_{k=1}^K \frac{\gamma_{k-1}^{r-1}}{2}(D(x_k, y) - D(x_{k+1}, y)) + 0.5 \sum_{k=1}^K (\gamma_k^{r-1} - \gamma_{k-1}^{r-1}) p_{\min}^{-1} M \\ &\quad + 0.5 \sum_{k=1}^K p_{i_k}^{-1} \gamma_k^{1+r} \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 + \sum_{k=1}^K \gamma_k^r(y - x_k)^T (w_k + \Delta_k). \end{aligned}$$

This yields

$$\begin{aligned} \sum_{k=1}^K \gamma_k^r(f(x_k) - f(y)) &\leq \frac{\gamma_0^{r-1}}{2}(D(x_1, y) - D(x_{K+1}, y)) + 0.5(\gamma_K^{r-1} - \gamma_0^{r-1}) p_{\min}^{-1} M \\ &\quad + 0.5 \sum_{k=1}^K p_{i_k}^{-1} \gamma_k^{1+r} \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 + \sum_{k=1}^K \gamma_k^r(y - x_k)^T (w_k + \Delta_k). \end{aligned} \quad (2)$$

Consider equation (2). Letting $k = 0$ we have

$$\gamma_0^r(f(x_0) - f(y)) \leq 0.5\gamma_0^{r-1}(D(x_0, y) - D(x_1, y)) + 0.5p_{i_0}^{-1} \gamma_0^{1+r} \|\nabla_{x^{(i_0)}} F(x_0, \xi_0)\|^2 + \gamma_0^r(y - x_0)^T (w_0 + \Delta_0). \quad (3)$$

Adding (2) and (3), we obtain

$$\begin{aligned} \sum_{k=0}^K \gamma_k^r f(x_k) - \sum_{k=0}^K \gamma_k^r f(y) &\leq \frac{\gamma_0^{r-1}}{2}(D(x_0, y) - D(x_{K+1}, y)) + 0.5(\gamma_K^{r-1} - \gamma_0^{r-1}) p_{\min}^{-1} M \\ &\quad + 0.5 \sum_{k=0}^K p_{i_k}^{-1} \gamma_k^{1+r} \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 + \sum_{k=0}^K \gamma_k^r(y - x_k)^T (w_k + \Delta_k). \end{aligned}$$

Dropping $D(x_{K+1}, y)$ and dividing both sides by $\sum_{t=0}^K \gamma_t^r$ we have

$$\sum_{k=0}^K \left(\frac{\gamma_k^r}{\sum_{t=0}^K \gamma_t^r} (f(x_k)) \right) - f(y) \leq \left(\sum_{t=0}^K \gamma_t^r \right)^{-1} \left(\frac{\gamma_0^{r-1}}{2} D(x_0, y) + 0.5 (\gamma_K^{r-1} - \gamma_0^{r-1}) p_{\min}^{-1} M \right. \\ \left. + 0.5 \sum_{k=0}^K p_{i_k}^{-1} \gamma_k^{1+r} \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 + \sum_{k=0}^K \gamma_k^r (y - x_k)^T (w_k + \Delta_k) \right). \quad (4)$$

In-class assignment 4:

Consider the definition of \bar{x}_K . Show that we have

$$f(\bar{x}_K) \leq \sum_{k=0}^K \left(\frac{\gamma_k^r}{\sum_{t=0}^K \gamma_t^r} (f(x_k)) \right).$$

Using the definition of M , and invoking the aforementioned result, we have

$$f(\bar{x}_K) - f(y) \leq \left(\sum_{t=0}^K \gamma_t^r \right)^{-1} \left(\frac{\gamma_0^{r-1}}{2} p_{\min}^{-1} M + 0.5 (\gamma_K^{r-1} - \gamma_0^{r-1}) p_{\min}^{-1} M \right. \\ \left. + 0.5 \sum_{k=0}^K p_{i_k}^{-1} \gamma_k^{1+r} \|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2 + \sum_{k=0}^K \gamma_k^r (y - x_k)^T (w_k + \Delta_k) \right).$$

Thus we obtain

$$f(\bar{x}_K) - f(y) \leq \left(\sum_{t=0}^K \gamma_t^r \right)^{-1} \left(0.5 p_{\min}^{-1} M \gamma_K^{r-1} + 0.5 \sum_{k=0}^K p_{i_k}^{-1} \gamma_k^{1+r} \mathbb{E} [\|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2] + \sum_{k=0}^K \gamma_k^r \mathbb{E} [(y - x_k)^T (w_k + \Delta_k)] \right).$$

Next, taking expectation on both sides we obtain

$$\mathbb{E} [f(\bar{x}_K)] - f(y) \leq \left(\sum_{t=0}^K \gamma_t^r \right)^{-1} \left(0.5 p_{\min}^{-1} M \gamma_K^{r-1} + 0.5 \sum_{k=0}^K p_{i_k}^{-1} \gamma_k^{1+r} \mathbb{E} [\|\nabla_{x^{(i_k)}} F(x_k, \xi_k)\|^2] + \sum_{k=0}^K \gamma_k^r \mathbb{E} [(y - x_k)^T (w_k + \Delta_k)] \right).$$

Invoking Lemma 2 we obtain

$$\mathbb{E} [f(\bar{x}_K)] - f(y) \leq \left(\sum_{t=0}^K \gamma_t^r \right)^{-1} \left(0.5 p_{\min}^{-1} M \gamma_K^{r-1} + 0.5 \sum_{k=0}^K p_{i_k}^{-1} \gamma_k^{1+r} p_{\min}^{-1} (\sigma^2 + C^2) \right).$$

(b) From part (a), substituting the step size and $y := x^*$ and $r := 0$ we obtain

$$\mathbb{E} [f(\bar{x}_K)] - f^* \leq \frac{0.5 p_{\min}^{-1} M \gamma_0^{-1} \sqrt{K+1} + 0.5 p_{\min}^{-1} (\sigma^2 + C^2) \sum_{k=0}^K \gamma_k}{K+1}.$$

It can be shown that

$$\sum_{k=0}^K \gamma_k \leq 2\gamma_0 \sqrt{K+1}, \quad \text{for all } K \geq 3.$$

Thus, we obtain

$$\mathbb{E} [f(\bar{x}_K)] - f^* \leq \frac{0.5 p_{\min}^{-1} M \gamma_0^{-1} \sqrt{K+1} + \gamma_0 p_{\min}^{-1} (C^2 + \sigma^2) \sqrt{K+1}}{K+1}.$$

This implies the desired bound.

Stochastic Optimization (Module 7: Two-stage Stochastic Programming)

Module's main topics/objectives:

- What is stochastic programming?
- What are the standard approaches in addressing stochastic programs?
- How to solve two-stage stochastic programs using zeroth-order SGD?

Stochastic programming

Consider a stochastic optimization problem of the form

$$\begin{aligned} & \text{minimize}_x \quad h(x) + \mathbb{E}_\xi \left[\min_{y \in Y(x, \xi)} q(y, \xi) \right] \\ & \text{subject to:} \\ & \quad x \in X. \end{aligned} \quad (\text{P})$$

- $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ are decision variables
- $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is a deterministic cost function.
- $q : \mathbb{R}^m \times \Omega \rightarrow \mathbb{R}$ is a stochastic cost function.
- $\xi : \Omega \rightarrow \mathbb{R}^d$ is a d -dimensional random variable.
- $Y : \mathbb{R}^m \times \Omega \rightarrow \mathbb{R}$ is a constraint set for variable y .
- $\mathbb{E}[\bullet]$ denotes the expectation operator with respect to ξ .

Problem reformulation as a two-stage problem

We consider solving the following [two-stage stochastic programming](#) problem.

$$\begin{aligned} & \text{minimize}_x \quad h(x) + \mathbb{E}[Q(x, \xi)] \\ & \text{subject to:} \\ & \quad x \in X. \end{aligned} \quad (\text{P}^{(1s)})$$

where $Q(x, \xi)$ is called **value function** that is the optimal objective of the following problem

$$\begin{aligned} Q(x, \xi) & \triangleq \min_y \quad q(y, \xi) \\ & \text{s. t.} \\ & \quad y \in Y(x, \xi). \end{aligned} \quad (\text{P}_\xi^{(2s)})$$

x : First stage decision

y : Second stage decision

$h(x)$: First stage cost

$\mathbb{E}[Q(x, \xi)]$: Second stage cost

Note that the optimal objective function of $\text{P}_\xi^{(2s)}$ is characterized by x and ξ . This is because any optimal solution of this problem will be in terms of x and ξ . Let us denote an optimal solution as $y^* = y(x, \xi)$, where $y(x, \xi)$ denotes a mapping (possibly unknown). Then,

$$q(y^*, \xi) = q(y(x, \xi), \xi).$$

The value function $Q(x, \xi)$ is indeed an implicit representation of $q(y(x, \xi), \xi)$.

In applications, often $Q(x, \xi)$ is unknown to us.

Example: The news vendor problem

x : A news vendor goes to the publisher every morning and buys x newspaper

c : unit purchase price from the publisher

u : an upper bound on x , due to the news vendor's purchase power

The news vendor then walks along the streets to sell as many newspapers as possible.

q : unit selling price

r : value of unsold newspaper returned to publisher where $r < c$

ξ : demand for the newspaper per day

Decision to make: How many newspaper to buy from the publisher?

Let us define

y_1 : effective sales

y_2 : number of newspapers returned at the end of the day

Then, the optimization problem is as follows.

$$\begin{aligned} & \text{minimize} && cx + \mathbb{E}[Q(x, \xi)] \\ & \text{subject to:} && \\ & && 0 \leq x \leq u. \end{aligned}$$

where $Q(x, \xi)$ is the optimal objective of the following problem

$$\begin{aligned} Q(x, \xi) &\triangleq \min_{y_1, y_2} - (qy_1 + ry_2) \\ &\text{s. t.} \\ & y_1 \leq \xi, \\ & y_2 \leq x - y_1, \\ & y_1 \geq 0, \\ & y_2 \geq 0. \end{aligned}$$

In-class assignment 1:

Write the following terms by reformulating the news vendor problem as a formal two-stage stochastic programming model given by (P) and (P_ξ^2) .

$$h(x) = cx$$

$$X = [0, u]$$

$$y = (y_1, y_2)$$

$$q(y, \xi) = - (qy_1 + ry_2)$$

$$Y(x, \xi) = \{(y_1, y_2) \in \mathbb{R}_+^2 \mid y_1 \leq \xi, \quad y_1 + y_2 \leq x\}$$

A linear program solve: `scipy.optimize.linprog`

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.linprog.html> (<https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.linprog.html>)

Let us assume that

$$c = 1.5$$

$$q = 2$$

$$r = 1$$

ξ is uniformly at random between 40 and 50

$$x = 30$$

$$u = 100$$

Now, consider a random realization of the demand.

```
In [363]: 1 import numpy as np
          2 ksi = np.random.randint(40, 51, size=1)[0]
          3 print(ksi)
          4 x = 45
```

47

```
In [364]: 1 from scipy.optimize import linprog
2
3 c_vector = [-2 , -1]
4
5 A = [[1,1]]
6 b = [x]
7
8 y1_bounds = (0, ksi)
9 y2_bounds = (0, None)
10
11 news_vendor_output = linprog(c_vector, A_ub=A, b_ub=b,
12                             bounds=[y1_bounds, y2_bounds],
13                             method='simplex')
```

```
In [365]: 1 print(news_vendor_output)

con: array([], dtype=float64)
fun: -90.0
message: 'Optimization terminated successfully.'
nit: 2
slack: array([0.])
status: 0
success: True
x: array([45., 0.])
```

```
In [366]: 1 news_vendor_output.fun
```

```
Out[366]: -90.0
```

In-class assignment 2:

Find the closed-form formula of $Q(x, \xi)$ for a given x and ξ .

$$y_1 = \min(x, \xi)$$

$$y_2 = x - y_1 = x - \min(x, \xi) = x + \max(-x, -\xi) = \max(0, x - \xi)$$

$$Q(x, \xi) = \min_y q(y, \xi) = - (q \min(x, \xi) + r \max(0, x - \xi)) = - (2 \min(x, \xi) + \max(0, x - \xi))$$

$$Q(45, 47) = -90$$

Reformulation of SP as stochastic optimization (SO)

$$\begin{array}{ll} \text{minimize}_x & \mathbb{E}[F(x, \xi)] \\ \text{subject to:} & x \in X. \end{array}$$

(SO)

$$F(x, \xi) \triangleq h(x) + Q(x, \xi).$$

or

$$F(x, \xi) \triangleq h(x) + \min_{y \in Y(x, \xi)} q(y, \xi).$$

$$x_{k+1} = \mathcal{P}_X(x_k - \gamma_k (\nabla h(x) + \nabla Q(x, \xi)))$$

Monte Carlo sampling schemes

1. Sample Average Approximation (SAA)

This is the most direct sampling approach. In this approach, problem (SO) is approximated by the following (SAA) problem for a sufficiently large sample size N

$$\begin{array}{ll} \text{minimize}_x & \frac{1}{N} \sum_{j=1}^N F(x, \xi_j) \\ \text{subject to:} & x \in X. \end{array}$$

Then, a deterministic optimization, such as gradient descent, is usually employed to solve P_{SAA} .

Employing this approach for (SP) results in the following approximate problem

$$\begin{array}{ll} \text{minimize}_x & h(x) + \frac{1}{N} \sum_{j=1}^N Q(x, \xi_j) \\ \text{subject to:} & x \in X. \end{array}$$

$$\text{where } Q(x, \xi_j) \triangleq \min_y q(y, \xi_j) \\ \text{s. t.} \\ y \in Y(x, \xi_j).$$

Under some assumptions (to be discussed later), we reach as the following problem

$$\begin{aligned} &\text{minimize}_x \quad h(x) + \min_{y_N(x)} \frac{1}{N} \sum_{j=1}^N q(y(x, \xi_j), \xi_j) \\ &\text{subject to:} \\ &\quad x \in X, \\ &\quad y(x, \xi_j) \in Y(x, \xi_j), \quad \text{for } j = 1, \dots, N. \end{aligned}$$

where $y_N(x) \triangleq (y(x, \xi_1); \dots, y(x, \xi_N))$

2. Stochastic Approximation (SA)

In this scheme, sampling is done as the computational method progresses.

The most popular example of an SA method is SGD and SQN discussed before.

Applying SGD to (SO), we have

Challenges in employng SGD for solving SP

- $F(x, \xi)$ may not be differentiable
- Even if $F(x, \xi)$ is differentiable, we may not know the formula of its gradient
- $F(x, \xi)$ may not be convex

This motivates the use of zeroth-order SGD for solving SP problems.

Algorithm outline (ZO-SGD)

Initialization: Choose a random vector $x_0 \in X$, a stepsize sequence $\{\gamma_k\}$, $K \geq 1$, a smoothing sequence $\{\eta_k\}$

for $k = 0, \dots, K - 1$

Generate random realization of ξ , denoted by ξ_k

Generate random realization of $v \in \eta_k \mathbb{S}$, denoted by v_k

Evaluate the zeroth-order stochastic gradient as follows.

$$g_{\eta_k}(x_k, v_k, \xi_k) = \left(\frac{n}{\eta_k} \right) (F(x_k + v_k, \xi_k) - F(x_k, \xi_k)) \frac{v_k}{\|v_k\|}.$$

Update the main iterate as follows.

$$x_{k+1} = \mathcal{P}_X(x_k - \gamma g_{\eta_k}(x_k, v_k, \xi_k)).$$

end for

Output: Return a suitable output (depends on convexity or nonconvexity of F)

Variance reduced zeorth-order methods for nonconvex implicit functions

Algorithm outline (VR-ZO-SGD for nondifferentiable nonconvex case)

Initialization: Choose a random vector $x_0 \in X$, a stepsize γ , $K \geq 1$, a random value R from $\{0, \dots, K - 1\}$,

for $k = 0, \dots, K - 1$

for $j = 1, \dots, N_k$

Generate random realization of ξ , denoted by $\xi_{j,k}$

Generate random realization of $v \in \eta\mathbb{S}$, denoted by $v_{j,k}$

Evaluate the zeroth-order stochastic gradient as follows.

$$g_\eta(x_k, v_{j,k}, \xi_{j,k}) = \left(\frac{n}{\eta} \right) (F(x_k + v_{j,k}, \xi_{j,k}) - F(x_k, \xi_{j,k})) \frac{v_{j,k}}{\|v_{j,k}\|}.$$

end for

Update the main iterate as follows.

$$x_{k+1} = \mathcal{P}_X \left(x_k - \gamma \frac{\sum_{j=1}^{N_k} g_\eta(x_k, v_{j,k}, \xi_{j,k})}{N_k} \right).$$

end for

Output: Return x_R

Lemma 1 (Properties of the zeroth-order stochastic gradient): Consider the VR-ZO-SGD method. Then, for all j and k we have:

- (i) $\mathbb{E} [g_\eta(x_k, v_{j,k}, \xi_{j,k}) \mid x_k] = \nabla f_\eta(x_k).$
- (ii) $\mathbb{E} [\|g_\eta(x_k, v_{j,k}, \xi_{j,k})\|^2 \mid x_k] \leq L_0^2 n^2.$

Proof: To show part (i), we can write

$$\begin{aligned} \mathbb{E} [g_\eta(x_k, v_{j,k}, \xi_{j,k}) \mid x_k] &= \mathbb{E} \left[\left(\frac{n}{\eta} \right) (F(x_k + v_{j,k}, \xi_{j,k}) - F(x_k, \xi_{j,k})) \frac{v_{j,k}}{\|v_{j,k}\|} \mid x_k \right] \\ &= \left(\frac{n}{\eta} \right) \mathbb{E}_{v_{j,k}} \left[\mathbb{E}_{\xi_{j,k}} \left[(F(x_k + v_{j,k}, \xi_{j,k}) - F(x_k, \xi_{j,k})) \frac{v_{j,k}}{\|v_{j,k}\|} \mid x_k \cup \{v_{j,k}\} \right] \right] \\ &= \left(\frac{n}{\eta} \right) \mathbb{E}_{v_{j,k}} \left[(f(x_k + v_{j,k}) - f(x_k)) \frac{v_{j,k}}{\|v_{j,k}\|} \mid x_k \right] \\ &\stackrel{\mathbb{E}[v_{j,k}]=0}{=} \left(\frac{n}{\eta} \right) \mathbb{E}_{v_{j,k}} \left[f(x_k + v_{j,k}) \frac{v_{j,k}}{\|v_{j,k}\|} \mid x_k \right] \\ &= \nabla f_\eta(x_k). \end{aligned}$$

To show part (ii), we can write

$$\begin{aligned} \mathbb{E} [\|g_\eta(x_k, v_{j,k}, \xi_{j,k})\|^2 \mid x_k] &= \left(\frac{n}{\eta} \right)^2 \mathbb{E} \left[\left\| (F(x_k + v_{j,k}, \xi_{j,k}) - F(x_k, \xi_{j,k})) \frac{v_{j,k}}{\|v_{j,k}\|} \right\|^2 \mid x_k \right] \\ &\stackrel{\text{Law of total expectation}}{=} \left(\frac{n}{\eta} \right)^2 \mathbb{E}_{v_{j,k}} \left[\mathbb{E}_{\xi_{j,k}} \left[\left\| (F(x_k + v_{j,k}, \xi_{j,k}) - F(x_k, \xi_{j,k})) \frac{v_{j,k}}{\|v_{j,k}\|} \right\|^2 \mid x_k \cup \{v_{j,k}\} \right] \right] \\ &= \left(\frac{n}{\eta} \right)^2 \mathbb{E}_{v_{j,k}} \left[\mathbb{E}_{\xi_{j,k}} \left[\|(F(x_k + v_{j,k}, \xi_{j,k}) - F(x_k, \xi_{j,k}))\|^2 \mid x_k \cup \{v_{j,k}\} \right] \right] \\ &\leq \left(\frac{n}{\eta} \right)^2 \mathbb{E}_{v_{j,k}} \left[\mathbb{E}_{\xi_{j,k}} \left[L_0^2 (\xi_{j,k})^2 \|v_{j,k}\|^2 \mid x_k \cup \{v_{j,k}\} \right] \right] \\ &= \left(\frac{n}{\eta} \right)^2 \mathbb{E}_{v_{j,k}} \left[L_0^2 \|v_{j,k}\|^2 \mid x_k \right] \\ &\stackrel{v_{j,k} \in \eta\mathbb{S}}{=} \left(\frac{n}{\eta} \right)^2 \mathbb{E}_{v_{j,k}} \left[L_0^2 \eta^2 \mid x_k \right] \\ &= L_0^2 n^2. \end{aligned}$$

Let us define

$$g_\eta^{N_k}(x_k) \triangleq \frac{\sum_{j=1}^{N_k} g_\eta(x_k, v_{j,k}, \xi_{j,k})}{N_k}.$$

Also,

$$w_k \triangleq g_{\eta}^{N_k}(x_k) - \nabla f_{\eta}(x_k).$$

and

$$w_{j,k} \triangleq g_{\eta}(x_k, v_{j,k}, \xi_{j,k}) - \nabla f_{\eta}(x_k).$$

Corollary 1: For any j, k we have

$$(i) \mathbb{E} [w_{j,k} \mid x_k] = 0.$$

$$(ii) \mathbb{E} [\|w_{j,k}\|^2 \mid x_k] \leq L_0^2 n^2.$$

Proof:

$$(i) \mathbb{E} [g_{\eta}(x_k, v_{j,k}, \xi_{j,k}) - \nabla f_{\eta}(x_k) \mid x_k] = \mathbb{E} [g_{\eta}(x_k, v_{j,k}, \xi_{j,k}) \mid x_k] - \nabla f_{\eta}(x_k) \stackrel{\text{Lemma 1(i)}}{=} 0.$$

(ii) **In-class assignment 3:** Show (ii)

See Module 5

Lemma 2 (Properties of the mini-batch zeroth-order stochastic gradient): Consider the VR-ZO-SGD method. Then, for all k we have:

$$(i) \mathbb{E} [w_k \mid x_k] = 0.$$

$$(ii) \mathbb{E} [\|w_k\|^2 \mid x_k] \leq \frac{L_0^2 n^2}{N_k}.$$

Proof:

(i) From the definition of $g_{\eta}^{N_k}(x_k)$, we can write

$$\mathbb{E} [g_{\eta}^{N_k}(x_k) \mid x_k] = \mathbb{E} \left[\frac{\sum_{j=1}^{N_k} g_{\eta}(x_k, v_{j,k}, \xi_{j,k})}{N_k} \mid x_k \right] = \frac{\sum_{j=1}^{N_k} \mathbb{E} [g_{\eta}(x_k, v_{j,k}, \xi_{j,k}) \mid x_k]}{N_k} \stackrel{\text{Lemma 1(i)}}{=} \frac{\sum_{j=1}^{N_k} \nabla f_{\eta}(x_k)}{N_k} = \frac{N_k \nabla f_{\eta}(x_k)}{N_k} = \nabla f_{\eta}(x_k).$$

(ii) We can write

$$\begin{aligned} \mathbb{E} [\|w_k\|^2 \mid x_k] &= \mathbb{E} \left[\left\| \frac{\sum_{j=1}^{N_k} g_{\eta}(x_k, v_{j,k}, \xi_{j,k})}{N_k} - \nabla f_{\eta}(x_k) \right\|^2 \mid x_k \right] \\ &= \mathbb{E} \left[\left\| \frac{\sum_{j=1}^{N_k} (g_{\eta}(x_k, v_{j,k}, \xi_{j,k}) - \nabla f_{\eta}(x_k))}{N_k} \right\|^2 \mid x_k \right] \\ &= \frac{\mathbb{E} [\left\| \sum_{j=1}^{N_k} (g_{\eta}(x_k, v_{j,k}, \xi_{j,k}) - \nabla f_{\eta}(x_k)) \right\|^2 \mid x_k]}{N_k^2} \\ &= \frac{\mathbb{E} [\left\| \sum_{j=1}^{N_k} w_{j,k} \right\|^2 \mid x_k]}{N_k^2} \\ &= \frac{\mathbb{E} [\sum_{j=1}^{N_k} \|w_{j,k}\|^2 + 2 \sum_{j=1}^{N_k} \sum_{i>j}^{N_k} w_{i,k}^T w_{j,k} \mid x_k]}{N_k^2} \end{aligned}$$

In-class assignment 5:

Show that for any $i \neq j, k$, we have $\mathbb{E} [w_{i,k}^T w_{j,k} \mid x_k] = 0$.

Hint: $\mathbb{E} [w_{j,k} \mid x_k] = 0$.

Thus, we obtain

$$\mathbb{E} [\|w_k\|^2 \mid x_k] = \frac{\mathbb{E} [\sum_{j=1}^{N_k} \|w_{j,k}\|^2]}{N_k^2} = \frac{\sum_{j=1}^{N_k} \mathbb{E} [\|w_{j,k}\|^2]}{N_k^2} \stackrel{\text{Corollary 1(ii)}}{\leq} \frac{\sum_{j=1}^{N_k} L_0^2 n^2}{N_k^2} = \frac{N_k L_0^2 n^2}{N_k^2} = \frac{L_0^2 n^2}{N_k}.$$

$$\|u_1 + u_2 + \dots + u_N\|^2 \leq N \sum_{i=1}^N \|u_i\|^2$$

Corollary 2: Consider the ZO-SGD method. Then, for all k we have:

$$(i) \mathbb{E} [w_k] = 0.$$

$$(ii) \mathbb{E} [\|w_k\|^2] \leq \frac{L_0^2 n^2}{N_k}.$$

Proof: The proof follows by employing the law of total expectation on Lemma 2.

Convergence analysis of VR-ZO-SGD (nonconvex case)

Theorem 1: Let $f^* := \min_{x \in \mathbb{R}^n} f(x)$ and let Assumptions 1,2, and 3 hold (See Module 5 for these assumptions). Suppose x_R is generated by the VR-ZO-SGD method. Also, choose γ and η such that $\frac{\gamma n L_0}{\eta} < 0.5$. Let $N_k := N = \sqrt{K}$ for all k . Then, for all $K > \frac{4n^2 L_0^2}{\eta^2}$ we have

$$\mathbb{E} \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq \frac{8 \left(\mathbb{E} [f(x_0)] - f^* + 2L_0\eta + L_0^2 n^2 \right)}{\sqrt{K}}.$$

Proof:

Note that from Module 5, the function f_η is $\frac{nL_0}{\eta}$ -smooth. Invoking the descent lemma for the smoothed function $f_\eta(x)$ we have

$$f_\eta(x_{k+1}) \leq f_\eta(x_k) + \nabla f_\eta(x_k)^T (x_{k+1} - x_k) + \frac{nL_0}{2\eta} \|x_{k+1} - x_k\|^2.$$

Invoking the projection theorem (see Lecture_notes_20230127 by choosing $\hat{x} := x_k - \gamma(\nabla f_\eta(x_k) + w_k)$), we obtain

$$(x_{k+1} - (x_k - \gamma(\nabla f_\eta(x_k) + w_k)))^T (x - x_{k+1}) \geq 0, \quad \forall x \in X.$$

Let us choose $x := x_k$. Then we have

$$(x_{k+1} - x_k + \gamma(\nabla f_\eta(x_k) + w_k))^T (x_k - x_{k+1}) \geq 0.$$

Thus we have

$$-\|x_k - x_{k+1}\|^2 + \gamma w_k^T (x_k - x_{k+1}) \geq \nabla f_\eta(x_k)^T (x_{k+1} - x_k).$$

Rearranging the terms we obtain

$$\nabla f_\eta(x_k)^T (x_{k+1} - x_k) \leq -\frac{1}{\gamma} \|x_k - x_{k+1}\|^2 + w_k^T (x_k - x_{k+1}).$$

From the preceding relation and the first inequality we obtain

$$f_\eta(x_{k+1}) \leq f_\eta(x_k) - \frac{1}{\gamma} \|x_{k+1} - x_k\|^2 + w_k^T (x_k - x_{k+1}) + \frac{nL_0}{2\eta} \|x_{k+1} - x_k\|^2.$$

We can write $w_k^T (x_k - x_{k+1}) \leq \frac{\gamma}{2} \|w_k\|^2 + \frac{1}{2\gamma} \|x_{k+1} - x_k\|^2$. This implies that

$$\begin{aligned} f_\eta(x_{k+1}) &\leq f_\eta(x_k) - \frac{1}{2\gamma} \|x_{k+1} - x_k\|^2 + \frac{\gamma}{2} \|w_k\|^2 + \frac{nL_0}{2\eta} \|x_{k+1} - x_k\|^2 \\ &= f_\eta(x_k) - \frac{1}{2\gamma} \left(1 - \frac{\gamma n L_0}{\eta} \right) \|x_{k+1} - x_k\|^2 + \frac{\gamma}{2} \|w_k\|^2. \end{aligned}$$

Invoking Remark 1, we obtain

$$f_\eta(x_{k+1}) \leq f_\eta(x_k) - \frac{1}{2\gamma} \left(1 - \frac{\gamma n L_0}{\eta} \right) \gamma^2 \|\tilde{G}_{\eta, \frac{1}{\gamma}}(x_k)\|^2 + \frac{\gamma}{2} \|w_k\|^2.$$

Suppose $\frac{\gamma n L_0}{\eta} < 1$. Invoking Lemma 4, we get

$$\begin{aligned} f_\eta(x_{k+1}) &\leq f_\eta(x_k) - \frac{1}{2\gamma} \left(1 - \frac{\gamma n L_0}{\eta} \right) \gamma^2 (0.5) \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 + \frac{1}{2\gamma} \left(1 - \frac{\gamma n L_0}{\eta} \right) \gamma^2 \|w_k\|^2 + \frac{\gamma}{2} \|w_k\|^2 \\ &\leq f_\eta(x_k) - \frac{\gamma}{4} \left(1 - \frac{\gamma n L_0}{\eta} \right) \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 + \gamma \|w_k\|^2. \end{aligned}$$

Suppose $\frac{\gamma n L_0}{\eta} < 0.5$. Then, $\left(1 - \frac{\gamma n L_0}{\eta} \right) > 0.5$, implying that

$$f_\eta(x_{k+1}) \leq f_\eta(x_k) - \frac{\gamma}{8} \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 + \gamma \|w_k\|^2.$$

Thus we have

$$\frac{\gamma}{8} \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 \leq f_\eta(x_k) - f_\eta(x_{k+1}) + \gamma \|w_k\|^2.$$

Taking sum on both sides for $k = 0, \dots, K-1$ we obtain

$$\frac{\gamma}{8} \sum_{k=0}^{K-1} \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 \leq f_\eta(x_0) - f_\eta(x_K) + \gamma \sum_{k=0}^{K-1} \|w_k\|^2.$$

Also, note that from the definition of x_R we can write

$$\sum_{k=0}^{K-1} \|G_{\eta, \frac{1}{\gamma}}(x_k)\|^2 = (K - k_0) \mathbb{E}_R \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right].$$

We obtain

$$\frac{\gamma K}{8} \mathbb{E}_R \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq f_\eta(x_0) - f_\eta(x_K) + \gamma \sum_{k=0}^{K-1} \|w_k\|^2.$$

Invoking Lemma 1(iii) we have

$$|f_\eta(x) - f(x)| \leq L_0 \eta \quad \Rightarrow \quad f(x) - L_0 \eta \leq f_\eta(x) \leq f(x) + L_0 \eta.$$

We obtain

$$\frac{\gamma K}{8} \mathbb{E}_R \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq f(x_0) - f(x_K) + 2L_0 \eta + \gamma \sum_{k=0}^{K-1} \|w_k\|^2.$$

We have

$$\mathbb{E}_R \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq \frac{8(f(x_0) - f^* + 2L_0 \eta)}{\gamma K} + \frac{8 \sum_{k=0}^{K-1} \|w_k\|^2}{K}.$$

Taking expectation on both sides, we have

$$\mathbb{E} \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq \frac{8(\mathbb{E}[f(x_0)] - f^* + 2L_0 \eta)}{\gamma K} + \frac{8 \sum_{k=0}^{K-1} \mathbb{E}[\|w_k\|^2]}{K}.$$

Invoking Corollary 2 we obtain

$$\mathbb{E} \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq \frac{8(\mathbb{E}[f(x_0)] - f^* + 2L_0 \eta)}{\gamma K} + \frac{8 \sum_{k=0}^{K-1} \frac{L_0^2 n^2}{N_k}}{K}.$$

Let $N_k := N$ for all k . Then

$$\mathbb{E} \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq \frac{8(\mathbb{E}[f(x_0)] - f^* + 2L_0 \eta)}{\gamma K} + \frac{8L_0^2 n^2}{N}.$$

Let $N = \sqrt{K}$ and $\gamma = \frac{1}{\sqrt{K}}$. Note that in view of $\frac{\gamma n L_0}{\eta} < 0.5$, we must have $K > \frac{4n^2 L_0^2}{\eta^2}$. Then

$$\mathbb{E} \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] \leq \frac{8(\mathbb{E}[f(x_0)] - f^* + 2L_0 \eta)}{\sqrt{K}} + \frac{8L_0^2 n^2}{\sqrt{K}}.$$

Iteration and sample complexity

Let us assume that we want to choose the maximum iteration number K and the sample size N such that we have

$$\mathbb{E} \left[\|G_{\eta, \frac{1}{\gamma}}(x_R)\|^2 \right] < \epsilon,$$

where ϵ is a desired accuracy level.

How large K and N should be chosen to meet this accuracy level?

Let us choose $N = \sqrt{K}$ and K such that

$$\frac{8(\mathbb{E}[f(x_0)] - f^* + 2L_0 \eta + L_0^2 n^2)}{\sqrt{K}} \leq \epsilon.$$

This is satisfied if we choose K such that

$$K \geq \frac{64(\mathbb{E}[f(x_0)] - f^* + 2L_0 \eta + L_0^2 n^2)^2}{\epsilon^2}.$$

But note that from Theorem 1 we must have $K > \frac{4n^2 L_0^2}{\eta^2}$. Thus

$$K \geq \max \left\{ \frac{64(\mathbb{E}[f(x_0)] - f^* + 2L_0 \eta + L_0^2 n^2)^2}{\epsilon^2}, \frac{4n^2 L_0^2}{\eta^2} \right\} = \mathcal{O} \left(\max \left\{ \frac{L_0^4 n^4}{\epsilon^2}, \frac{L_0^2 n^2}{\eta^2} \right\} \right).$$

This is called iteration complexity. Also, the sample complexity is as follows.

$$N \geq \max \left\{ \frac{8(\mathbb{E}[f(x_0)] - f^* + 2L_0 \eta + L_0^2 n^2)}{\epsilon}, \frac{2nL_0}{\eta} \right\} = \mathcal{O} \left(\max \left\{ \frac{L_0^2 n^2}{\epsilon}, \frac{L_0 n}{\eta} \right\} \right).$$

Total sample complexity is $N \times K$

We say $g(N) = \mathcal{O}(N^\alpha)$ if there exists $M > 0$ such that $g(N) \leq MN^\alpha$

Stochastic Optimization (Module 8: Federated Learning)

Today's main topics:

- What is federated stochastic gradient descent (SGD) method (FedAvg) and Local SGD?
- How fast does Local SGD converge for nonconvex stochastic optimization?

Problem formulation

We consider solving the following **multi-agent stochastic optimization** problem.

$$\begin{aligned} &\text{minimize} && f(x) \triangleq \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\xi_i \in \mathcal{D}_i} [F_i(x, \xi_i)] \\ &\text{subject to:} && x \in \mathbb{R}^n. \end{aligned}$$

(P)

where

- \mathcal{D}_i denotes the local dataset associated with agent i , for $i \in [m] \triangleq \{1, \dots, m\}$
- $F_i : \mathbb{R}^n \times \mathcal{D}_i \rightarrow \mathbb{R}$ denotes the local stochastic loss function,

Throughout, we let $f_i(x) \triangleq \mathbb{E}_{\xi_i \in \mathcal{D}_i} [F_i(x, \xi_i)]$ denote the local objective function of agent $i \in [m]$.

Agent: An entity (e.g., mobile phone, sensor, computer, self-driving car, ...) that has a limited memory and computational power.

$x_{i,k} \in \mathbb{R}^n$ denotes the local copy of x maintained by agent i at time k

The goal is to develop a decentralized scheme where $x_{i,k}$ is updated using a variant of SGD by agent i and eventually (when $k \rightarrow \infty$), all $x_{i,k}$ s converge to x^* .

Challenges: <https://arxiv.org/pdf/1602.05629.pdf> (<https://arxiv.org/pdf/1602.05629.pdf>).

Data and device heterogeneity and communication among the agents are among the key challenges of FL. In more details, challenges are as follows:

- **Data privacy:** Each agent privately stores the local data.
- **Non-IID:** Often, agent's local dataset is not representative of the population distribution.
- **Unbalanced local datasets:** The local datasets maintained by the agents may have different sizes.
- **Massively distributed:** The number of agents may be extremely large.
- **Limited communication:** Agents (mobile devices) may be offline or afflicted by slow connections.

Algorithm outline (a simple variant of FedAvg)

Initialization: Server chooses a random initial point $\hat{x}_0 \in \mathbb{R}^n$, stepsize γ , synchronization indices $T_0 := 0$ and $T_r \geq 1$, where $r \geq 1$ denotes the communication round index

for $r = 0, \dots, R - 1$

Server broadcasts \hat{x}_r to all agents $x_{i,T_r} := \hat{x}_r, \quad \forall i \in [m]$.

for $k = T_r, \dots, T_{r+1} - 1$ in parallel by all agents $i \in [m]$

Agent i generates the random replicate $\xi_{i,k} \in \mathcal{D}_i$ locally

Agent i does a local update as $x_{i,k+1} := x_{i,k} - \gamma \nabla F_i(x_{i,k}, \xi_{i,k})$.

Lemma 1: Consider Definition 1. Under Assumption 1 we have for all $k \geq 0$ and all $i \in [m]$:

$$\begin{aligned}\mathbb{E}[w_{i,k} \mid \mathcal{F}_k] &= \mathbf{0}_n \\ \mathbb{E}[\|w_{i,k}\|^2 \mid \mathcal{F}_k] &\leq \sigma^2.\end{aligned}$$

Lemma 2: From the probability law, we have that $\mathbb{E}[\mathbb{E}[\bullet \mid \mathcal{F}_k]] = \mathbb{E}[\bullet]$.

Notation and definitions:

We use the notation

$$g_{i,k} \triangleq \nabla F(x_{i,k}, \xi_{i,k}).$$

Also, we define the auxiliary averaged iterate \bar{x}_k as

$$\bar{x}_k \triangleq \frac{1}{m} \sum_{i=1}^m x_{i,k}, \quad \text{for all } k \geq 0.$$

We define the average squared **consensus violation** denoted by e_k as

$$\bar{e}_k \triangleq \frac{1}{m} \sum_{i=1}^m \|x_{i,k} - \bar{x}_k\|^2, \quad \text{for all } k \geq 0.$$

We also define the averaged stochastic gradient mapping as

$$\bar{g}_k \triangleq \frac{1}{m} \sum_{i=1}^m g_{i,k}, \quad \text{for all } k \geq 0.$$

$$\bar{g}_k = \frac{1}{m} \sum_{i=1}^m \nabla F_i(x_{i,k}, \xi_{i,k}) \neq \frac{1}{m} \sum_{i=1}^m \nabla F_i(\bar{x}_k, \xi_{i,k})$$

$$f(x) \triangleq \frac{1}{m} \sum_{i=1}^m f_i(x)$$

$$\nabla f(x) = \frac{1}{m} \sum_{i=1}^m \nabla f_i(x)$$

$$\nabla f(\bar{x}_k) \neq \frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,k})$$

$$\frac{1}{m} \sum_{i=1}^m \nabla F_i(x_{i,k}, \xi_{i,k})$$

Lemma 1: Consider Local SGD. Then, for all $k \geq 0$ we have

$$\boxed{\bar{x}_{k+1} = \bar{x}_k - \gamma \bar{g}_k.}$$

Remark 1 (Compact representation of Local SGD):

The following equation, for $k \geq 0$, compactly represents the update rules of Local SGD.

$$x_{i,k+1} := \begin{cases} \frac{1}{m} \sum_{j=1}^m (x_{j,k} - \gamma g_{j,k}), & k \in \mathcal{I} \\ x_{i,k} - \gamma g_{i,k}, & k \notin \mathcal{I}. \end{cases}$$

In-class assignment 1: How are FedAvg and Local SGD different from each other?

Proof of Lemma 1:

Case 1: When $k \in \mathcal{I}$, from Remark 1 we can write

$$\begin{aligned}x_{i,k+1} &= \frac{1}{m} \sum_{j=1}^m (x_{j,k} - \gamma g_{j,k}) \\ &= \frac{1}{m} \sum_{j=1}^m x_{j,k} - \gamma \frac{1}{m} \sum_{j=1}^m g_{j,k} = \bar{x}_k - \gamma \bar{g}_k,\end{aligned}$$

where the last equation is implied by the definition of \bar{x}_k and \bar{g}_k . Taking an average on the both sides over $i = 1, \dots, m$, we obtain

$$\bar{x}_{k+1} = \bar{x}_k - \gamma \bar{g}_k.$$

Case 2: When $k \notin \mathcal{I}$, from Remark 1 we can write

$$x_{i,k+1} = x_{i,k} - \gamma g_{i,k}.$$

Summing over $i = 1, \dots, m$ on both sides and then dividing both sides by m , we obtain

$$\frac{1}{m} \sum_{i=1}^m x_{i,k+1} = \frac{1}{m} \sum_{i=1}^m x_{i,k} - \gamma \frac{1}{m} \sum_{i=1}^m g_{i,k}.$$

This implies the desired result.

The smooth nonconvex case

Assumption 2: Let each local function $f_i(x)$ be L -smooth.

suppose each f_i is L_i -smooth. Then, can we find L such that Assumption 2 is met? Yes: $L := \max_{i=1, \dots, m} L_i$

Lemma 2: For all $k \geq 0$ we have

$$\mathbb{E} [\|\tilde{g}_k\|^2 \mid \mathcal{F}_k] \leq 2L^2 \bar{e}_k + 2\|\nabla f(\bar{x}_k)\|^2 + \frac{\sigma^2}{m}.$$

$$\mathbb{E} [\|\tilde{g}_k\|^2] \leq 2L^2 \mathbb{E} [\bar{e}_k] + 2\mathbb{E} [\|\nabla f(\bar{x}_k)\|^2] + \frac{\sigma^2}{m}.$$

Proof of Lemma 2:

$$\begin{aligned} \mathbb{E} [\|\tilde{g}_k\|^2 \mid \mathcal{F}_k] &= \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i=1}^m g_{i,k} \right\|^2 \mid \mathcal{F}_k \right] \\ &= \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i=1}^m (\nabla f_i(x_{i,k}) + w_{i,k}) \right\|^2 \mid \mathcal{F}_k \right] \\ &= \left\| \frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,k}) \right\|^2 + \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i=1}^m w_{i,k} \right\|^2 \mid \mathcal{F}_k \right] + 2\mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,k}) \right)^T \left(\frac{1}{m} \sum_{i=1}^m w_{i,k} \right) \mid \mathcal{F}_k \right]. \end{aligned}$$

Let us analyze the first, the second, and the third terms.

We can bound the first term as follows.

$$\begin{aligned} \left\| \frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,k}) \right\|^2 &= \left\| \frac{1}{m} \sum_{i=1}^m (\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k) + \nabla f_i(\bar{x}_k)) \right\|^2 \\ &= \left\| \left(\frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k) \right) + \nabla f(\bar{x}_k) \right\|^2 \\ &\leq 2 \left\| \left(\frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k) \right) \right\|^2 + 2\|\nabla f(\bar{x}_k)\|^2 \end{aligned}$$

Note that given vectors $y_i \in \mathbb{R}^n$ for $i \in [m]$, we have $\|\frac{1}{m} \sum_{i=1}^m y_i\|^2 \leq \frac{1}{m} \sum_{i=1}^m \|y_i\|^2$. Utilizing this inequality together with Assumption 1, we obtain

$$\begin{aligned} \left\| \frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,k}) \right\|^2 &\leq \frac{2}{m} \sum_{i=1}^m \|\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k)\|^2 + 2\|\nabla f(\bar{x}_k)\|^2 \\ &\leq \frac{2L^2}{m} \sum_{i=1}^m \|x_{i,k} - \bar{x}_k\|^2 + 2\|\nabla f(\bar{x}_k)\|^2 \\ &= 2L^2 \bar{e}_k + 2\|\nabla f(\bar{x}_k)\|^2 \end{aligned}$$

We can bound the second term as follows. Invoking Lemma 1 and utilizing the independence of the random variables, we have

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i=1}^m w_{i,k} \right\|^2 \mid \mathcal{F}_k \right] &= \frac{1}{m^2} \mathbb{E} \left[\left\| \sum_{i=1}^m w_{i,k} \right\|^2 \mid \mathcal{F}_k \right] \\ &= \frac{1}{m^2} \mathbb{E} \left[\sum_{i=1}^m \|w_{i,k}\|^2 + 2 \sum_{i=1}^m \sum_{j=i+1}^m w_{i,k}^T w_{j,k} \mid \mathcal{F}_k \right] \\ &= \frac{1}{m^2} \mathbb{E} \left[\sum_{i=1}^m \|w_{i,k}\|^2 \mid \mathcal{F}_k \right] \\ &= \frac{1}{m^2} \sum_{i=1}^m \mathbb{E} [\|w_{i,k}\|^2 \mid \mathcal{F}_k] \\ &\leq \frac{1}{m^2} \sum_{i=1}^m \sigma^2 \\ &= \frac{\sigma^2}{m}. \end{aligned}$$

We can bound the third term as follows. We can write

$$\begin{aligned} \mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,k}) \right)^T \left(\frac{1}{m} \sum_{i=1}^m w_{i,k} \right) \mid \mathcal{F}_k \right] &= \left(\frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,k}) \right)^T \mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m w_{i,k} \right) \mid \mathcal{F}_k \right] \\ &= \left(\frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,k}) \right)^T \left(\frac{1}{m} \sum_{i=1}^m \mathbb{E} [w_{i,k} \mid \mathcal{F}_k] \right) \\ &= 0. \end{aligned}$$

Thus, from the preceding inequalities we obtain

One approach is to add/subtract $\nabla f_i(\bar{x}_k)$

another approach is to add/subtract $\nabla f(x_{i,k})$

Lemma 3: For all $k \geq 0$ we have

$$\mathbb{E} [\nabla f(\bar{x}_k)^T \bar{g}_k \mid \mathcal{F}_k] \geq 0.5 \|\nabla f(\bar{x}_k)\|^2 - 0.5L^2 \bar{e}_k.$$

Proof of Lemma 3:

$$\begin{aligned} \mathbb{E} [\nabla f(\bar{x}_k)^T \bar{g}_k \mid \mathcal{F}_k] &= \mathbb{E} \left[\nabla f(\bar{x}_k)^T \left(\frac{1}{m} \sum_{i=1}^m g_{i,k} \right) \mid \mathcal{F}_k \right] \\ &= \nabla f(\bar{x}_k)^T \mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m g_{i,k} \right) \mid \mathcal{F}_k \right] \\ &= \nabla f(\bar{x}_k)^T \left(\frac{1}{m} \sum_{i=1}^m \mathbb{E} [g_{i,k} \mid \mathcal{F}_k] \right) \\ &= \nabla f(\bar{x}_k)^T \left(\frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,k}) \right) \\ &= \nabla f(\bar{x}_k)^T \left(\frac{1}{m} \sum_{i=1}^m (\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k) + \nabla f_i(\bar{x}_k)) \right) \\ &= \nabla f(\bar{x}_k)^T \left(\frac{1}{m} \sum_{i=1}^m (\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k)) + \nabla f(\bar{x}_k) \right) \\ &= \nabla f(\bar{x}_k)^T \left(\frac{1}{m} \sum_{i=1}^m (\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k)) \right) + \|\nabla f(\bar{x}_k)\|^2. \end{aligned}$$

Next we utilize the following inequality. For any $u, v \in \mathbb{R}^n$, we have $u^T v \geq -0.5\|u\|^2 - 0.5\|v\|^2$. We obtain

$$\mathbb{E} [\nabla f(\bar{x}_k)^T \bar{g}_k \mid \mathcal{F}_k] \geq -0.5\|\nabla f(\bar{x}_k)\|^2 - 0.5 \left\| \frac{1}{m} \sum_{i=1}^m (\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k)) \right\|^2 + \|\nabla f(\bar{x}_k)\|^2.$$

Note that given vectors $y_i \in \mathbb{R}^n$ for $i \in [m]$, we have $\left\| \frac{1}{m} \sum_{i=1}^m y_i \right\|^2 \leq \frac{1}{m} \sum_{i=1}^m \|y_i\|^2$. Utilizing this inequality we obtain

$$\begin{aligned} \mathbb{E} [\nabla f(\bar{x}_k)^T \bar{g}_k \mid \mathcal{F}_k] &\geq 0.5\|\nabla f(\bar{x}_k)\|^2 - \frac{0.5}{m} \sum_{i=1}^m \|\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k)\|^2 \\ &\geq 0.5\|\nabla f(\bar{x}_k)\|^2 - \frac{0.5L^2}{m} \sum_{i=1}^m \|x_{i,k} - \bar{x}_k\|^2 \\ &= 0.5\|\nabla f(\bar{x}_k)\|^2 - 0.5L^2 \bar{e}_k. \end{aligned}$$

Definition (Communication frequency bound):

Let $H > 0$ denote an upper bound on the communication frequency, i.e., $H \geq \max_{r=0,1,\dots} |T_{r+1} - T_r|$.

Throughout, we assume that H is bounded.

Assumption 4 (Bounded gradient dissimilarity): Assume that there exists $B_1 \geq 0$ and $B_2 > 0$ such that for all $x \in \mathbb{R}^n$ we have

$$\frac{1}{m} \sum_{i=1}^m \|\nabla f_i(x)\|^2 \leq B_1^2 + B_2^2 \|\nabla f(x)\|^2.$$

Lemma 4 (Bound on average second momenent of the local stochastic gradients): Let Assumption 1, 2, 3, and 4 hold. Then, for any $k \geq 0$, we have

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E} [\|g_{i,k}\|^2] \leq 2L^2 \mathbb{E} [\bar{e}_k] + 2B_1^2 + 2B_2^2 \mathbb{E} [\|\nabla f(\bar{x}_k)\|^2] + \sigma^2.$$

Proof of Lemma 4:

$$\begin{aligned}
\frac{1}{m} \sum_{i=1}^m \mathbb{E} [\|g_{i,k}\|^2 \mid \mathcal{F}_k] &\leq \frac{1}{m} \sum_{i=1}^m \mathbb{E} [\|\nabla f_i(x_{i,k}) + w_{i,k}\|^2 \mid \mathcal{F}_k] \\
&= \frac{1}{m} \sum_{i=1}^m \mathbb{E} [\|\nabla f_i(x_{i,k})\|^2 + \|w_{i,k}\|^2 + 2w_{i,k}^T \nabla f_i(x_{i,k}) \mid \mathcal{F}_k] \\
&= \frac{1}{m} \sum_{i=1}^m (\|\nabla f_i(x_{i,k})\|^2 + \mathbb{E} [\|w_{i,k}\|^2 \mid \mathcal{F}_k]) \\
&\leq \frac{1}{m} \sum_{i=1}^m (\|\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k) + \nabla f_i(\bar{x}_k)\|^2 + \sigma^2) \\
&= \frac{2}{m} \sum_{i=1}^m (\|\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k)\|^2 + \|\nabla f_i(\bar{x}_k)\|^2) + \sigma^2 \\
&= \frac{2L^2}{m} \sum_{i=1}^m \|x_{i,k} - \bar{x}_k\|^2 + \frac{2}{m} \sum_{i=1}^m \|\nabla f_i(\bar{x}_k)\|^2 + \sigma^2 \\
&\leq 2L^2 \bar{e}_k + 2B_1^2 + 2B_2^2 \|\nabla f(\bar{x}_k)\|^2 + \sigma^2.
\end{aligned}$$

Taking expectation with respect to \mathcal{F}_k on both sides, and invoking the total law of expectation, we obtain the bound.

Lemma 5 (Bound on average consensus violation):

Let Assumptions 1, 2, 3, and 4 hold and let $H \geq 1$ be given. Let $B_2 = 0$ and $\gamma \leq \frac{1}{H \max\{1, 2L^2\}q}$. Then, for some arbitrary $q \geq 2$ we have

$$\mathbb{E} [\bar{e}_k] \leq \frac{\sqrt{3} (2B_1^2 + \sigma^2)}{q^2}.$$

Proof of Lemma 5: For any i at any communication round $r > 0$, for all $T_r \leq k \leq T_{r+1} - 1$ we have

$$x_{i,k+1} = x_{i,k} - \gamma g_{i,k}.$$

Equivalently, we can write

$$x_{i,k} = x_{i,k-1} - \gamma g_{i,k-1}, \quad \text{for all } T_r + 1 \leq k \leq T_{r+1}.$$

This implies that

$$x_{i,k} = x_{i,T_r} - \gamma \sum_{t=T_r}^{k-1} g_{i,t}, \quad \text{for all } T_r + 1 \leq k \leq T_{r+1}.$$

Again, we have $\hat{x}_r = x_{i,T_r}$. From the definition of \bar{x}_k , we can write

$$\bar{x}_{T_r} = \hat{x}_r$$

This implies that $\bar{x}_{T_r} = x_{i,T_r}$ for all i and r . In view of Lemma 1, we have

$$\bar{x}_k = x_{i,T_r} - \gamma \sum_{t=T_r}^{k-1} \bar{g}_{i,t}, \quad \text{for all } T_r + 1 \leq k \leq T_{r+1}.$$

For all $T_r + 1 \leq k \leq T_{r+1}$, we have

$$\begin{aligned}
\mathbb{E} [\bar{e}_k] &= \frac{1}{m} \sum_{i=1}^m \mathbb{E} [\|x_{i,k} - \bar{x}_k\|^2] \\
&= \frac{1}{m} \sum_{i=1}^m \mathbb{E} \left[\left\| \gamma \sum_{t=T_r}^{k-1} g_{i,t} - \gamma \sum_{t=T_r}^{k-1} \bar{g}_t \right\|^2 \right] \\
&= \frac{\gamma^2}{m} \sum_{i=1}^m \mathbb{E} \left[\left\| \sum_{t=T_r}^{k-1} (g_{i,t} - \bar{g}_t) \right\|^2 \right] \\
&\leq \frac{\gamma^2(k - T_r)}{m} \sum_{i=1}^m \sum_{t=T_r}^{k-1} \mathbb{E} [\|g_{i,t} - \bar{g}_t\|^2] \\
&\leq \frac{\gamma^2 H}{m} \sum_{t=T_r}^{k-1} \sum_{i=1}^m \mathbb{E} [\|g_{i,t}\|^2 + \|\bar{g}_t\|^2 - 2g_{i,t}^T \bar{g}_t] \\
&\leq \gamma^2 H \sum_{t=T_r}^{k-1} \mathbb{E} \left[\left(\frac{1}{m} \sum_{i=1}^m \|g_{i,t}\|^2 \right) + \|\bar{g}_t\|^2 - 2\|\bar{g}_t\|^2 \right] \\
&\leq \gamma^2 H \sum_{t=T_r}^{k-1} \left(\frac{1}{m} \sum_{i=1}^m \mathbb{E} [\|g_{i,t}\|^2] \right) \\
&\leq \gamma^2 H \sum_{t=T_r}^{k-1} (2L^2 \mathbb{E} [\bar{e}_t] + 2B_1^2 + 2B_2^2 \mathbb{E} [\|\nabla f(\bar{x}_t)\|^2] + \sigma^2).
\end{aligned}$$

We utilize the following result to establish a nonrecursive bound for the average consensus violation.

Lemma: Suppose for $1 \leq t \leq H$, the nonnegative sequence $\{a_t\}$ and parameter θ satisfy a recursive relation of the form

$$a_t \leq H\gamma^2 \sum_{j=0}^{t-1} (\beta a_j + \theta),$$

where $0 < \gamma < \frac{1}{H \max\{1, \beta\} a}$ where $q \geq 2$. Then, we have

$$f(x) \leq f(y) + \nabla f(y)^T (x - y) + \frac{L}{2} \|x - y\|^2.$$

$$f(\bar{x}_{k+1}) \leq f(\bar{x}_k) + \nabla f(\bar{x}_k)^T (\bar{x}_{k+1} - \bar{x}_k) + \frac{L}{2} \|\bar{x}_{k+1} - \bar{x}_k\|^2.$$

$$\bar{x}_{k+1} - \bar{x}_k = -\gamma \bar{g}_k$$

$$f(\bar{x}_{k+1}) \leq f(\bar{x}_k) + \nabla f(\bar{x}_k)^T (-\gamma \bar{g}_k) + \frac{L}{2} \|\gamma \bar{g}_k\|^2.$$

$$f(\bar{x}_{k+1}) \leq f(\bar{x}_k) - \gamma \nabla f(\bar{x}_k)^T \bar{g}_k + \frac{L\gamma^2}{2} \|\bar{g}_k\|^2.$$

Taking expectations on both sides

$$\mathbb{E}[f(\bar{x}_{k+1})] \leq \mathbb{E}[f(\bar{x}_k)] - \gamma \mathbb{E}[\nabla f(\bar{x}_k)^T \bar{g}_k] + \frac{L\gamma^2}{2} \mathbb{E}[\|\bar{g}_k\|^2].$$

$$\mathbb{E}[f(\bar{x}_{k+1})] \leq \mathbb{E}[f(\bar{x}_k)] - \gamma \left(0.5 \mathbb{E}[\|\nabla f(\bar{x}_k)\|^2] - 0.5 L^2 \frac{\sqrt{3}(2B_1^2 + \sigma^2)}{q^2} \right) + \frac{L\gamma^2}{2} \left(2L^2 \frac{\sqrt{3}(2B_1^2 + \sigma^2)}{q^2} + 2 \mathbb{E}[\|\nabla f(\bar{x}_k)\|^2] + \frac{\sigma^2}{m} \right).$$

$$\gamma \leq \frac{1}{4L} \Rightarrow L\gamma^2 \leq \frac{\gamma}{4}$$

$$\gamma \leq \frac{1}{H \max\{1, 2L^2\} q}$$

$$\frac{\gamma}{4} \mathbb{E}[\|\nabla f(\bar{x}_k)\|^2] \leq \mathbb{E}[f(\bar{x}_k)] - \mathbb{E}[f(\bar{x}_{k+1})] + \frac{C_1 \gamma}{q^2} + \frac{C_2 \gamma^2}{m}$$

$$\mathbb{E}[\|\nabla f(\bar{x}_k)\|^2] \leq 4\gamma^{-1} (\mathbb{E}[f(\bar{x}_k)] - \mathbb{E}[f(\bar{x}_{k+1})]) + \frac{4C_1}{q^2} + \frac{4C_2 \gamma}{m}$$

$$\mathbb{E}[\|\nabla f(\bar{x}_{k^*})\|^2] = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(\bar{x}_k)\|^2] \leq 4 \frac{(\mathbb{E}[f(\bar{x}_0)] - \mathbb{E}[f(\bar{x}_K)])}{\gamma K} + \frac{4C_1}{q^2} + \frac{4C_2 \gamma}{m}$$

k^* is uniformly selected from $0, \dots, K-1$

$$\gamma := \frac{1}{\sqrt{K}}$$

$$q := \sqrt[4]{K}$$

$$\frac{1}{\sqrt{K}} \leq \frac{1}{H \max\{1, 2L^2\} \sqrt[4]{K}} \Rightarrow H \leq \frac{\sqrt[4]{K}}{\max\{1, 2L^2\}}$$

The optimal rate for FedAvg in nonconvex case $\frac{1}{\sqrt{mK}}$

$$\gamma := \frac{\sqrt{m}}{\sqrt{K}}$$

Stochastic Optimization (Module 9: DSGT)

Today's main topics:

- What is a distributed stochastic gradient tracking method over undirected networks?
- How fast does DSGT converge for strongly convex functions?
- Implementations on the MNIST dataset and comparison with SGD

Problem formulation

Consider solving the following unconstrained optimization problem of the form

$$\begin{aligned} &\text{minimize} \quad f(x) \triangleq \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\xi_i \in D_i} [f_i(x, \xi_i)] \\ &\text{subject to:} \\ &\quad x \in \mathbb{R}^n. \end{aligned} \tag{P}$$

- Here, we assume there are m agents.
- The function $f_i(\bullet, \xi_i) : \mathbb{R}^n \rightarrow \mathbb{R}$ is known only *locally*.
- The function $\mathbb{E} [f_i(x, \xi_i)]$ is possibly an unknown (or unavailable) deterministic function to agent i .
- Here, function $f_i(\bullet, \xi_i) : \mathbb{R}^n \rightarrow \mathbb{R}$ is a stochastic function and is *locally* known to the agent i .
- $\xi_i \in \mathbb{R}^d$ is a d -dimensional local random variable.
- $\mathbb{E}[\bullet]$ denotes the expectation operator with respect to ξ_i s.

Algorithm outline



- This is a distributed algorithm for optimization in **synchronous, static, and undirected** networks.
- Here, $[m]$ denotes the set of agents, i.e., $\{1, \dots, m\}$.

Notation:

$$\mathbf{x} := [x_1, x_2, \dots, x_m]^T, \quad \mathbf{y} := [y_1, y_2, \dots, y_m]^T \in \mathbb{R}^{m \times n},$$

$$\bar{\mathbf{x}} := \frac{1}{m} \mathbf{1}^T \mathbf{x} \in \mathbb{R}^{1 \times n}, \quad \bar{\mathbf{y}} := \frac{1}{m} \mathbf{1}^T \mathbf{y} \in \mathbb{R}^{1 \times n},$$

$$f(x) \triangleq \frac{1}{m} \sum_{i=1}^m f_i(x), \quad \mathbf{f}(\mathbf{x}) \triangleq \frac{1}{m} \sum_{i=1}^m f_i(x_i),$$

$$f_i(x) \triangleq \mathbb{E} [f_i(x, \xi_i) \mid x],$$

$$\boldsymbol{\xi} := [\xi_1, \xi_2, \dots, \xi_m]^T \in \mathbb{R}^{m \times d},$$

$$\mathbf{G}(\mathbf{x}, \boldsymbol{\xi}) \triangleq [\nabla f_1(x_1, \xi_1), \dots, \nabla f_m(x_m, \xi_m)]^T,$$

$$\mathbf{G}(\mathbf{x}) \triangleq \mathbb{E} [\mathbf{G}(\mathbf{x}, \boldsymbol{\xi}) \mid \mathbf{x}],$$

$$G(\mathbf{x}, \boldsymbol{\xi}) \triangleq \frac{1}{m} \mathbf{1}^T \mathbf{G}(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{m} \sum_{i=1}^m \nabla f_i(x_i, \xi_i),$$

$$G(\mathbf{x}) \triangleq \mathbb{E} [G(\mathbf{x}, \boldsymbol{\xi}) \mid \mathbf{x}] = \frac{1}{m} \sum_{i=1}^m \nabla f_i(x_i),$$

$$\bar{G}(x) \triangleq G(\mathbf{1}x^T) = \frac{1}{m} \sum_{i=1}^m \nabla f_i(x) = \nabla f(x).$$

History of the method: We define the filtration as follows:

$$\mathcal{F}_k \triangleq \cup_{i=1}^m \{x_{i,0}, \xi_{i,0}, \xi_{i,1}, \dots, \xi_{i,k-1}\} \quad \text{for all } k \geq 1,$$

$$\mathcal{F}_0 \triangleq \cup_{i=1}^m \{x_{i,0}\}.$$

Throughout, we assume $\|\bullet\|$ denotes the Euclidean norm of a vector and the Frobenius norm of a matrix.

Frobenius norm: For $\mathbf{A} \in \mathbb{R}^{m \times n}$, we Frobenius norm is defined as $\|\mathbf{A}\|_F \triangleq \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2}$.

Throughout, sometimes we use $\|\mathbf{A}\|$ to denote the Frobenius norm of matrix \mathbf{A} .

Definition 1 (Frobenius inner product):

Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{m \times n}$. Then,

$$\langle \mathbf{u}, \mathbf{v} \rangle \triangleq \text{Trace}(\mathbf{u}^T \mathbf{v}) = \sum_{i=1}^m \sum_{j=1}^n u_{ij} v_{ij}$$

Lemma 2: Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{m \times n}$. Then,

$$(a) \langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^m u_{i\bullet} v_{i\bullet}^T = \sum_{j=1}^n u_{\bullet j}^T v_{\bullet j}.$$

$$(b) \|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{v}\|^2, \text{ where } \|\bullet\| \text{ denotes the Frobenius norm of a matrix.}$$

$$(c) \text{ For any scalar } \lambda > 0, \text{ we have } |\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\| \leq \frac{1}{2} (\lambda \|\mathbf{u}\|^2 + \frac{1}{\lambda} \|\mathbf{v}\|^2).$$

$$(d) \|\mathbf{u}\|_F = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}.$$

$$(e) \langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$$

$$(f) \text{ Let } n = m. \text{ Then, the spectral norm of } \mathbf{u} \text{ (possibly non-symmetric) is equal to } \|\mathbf{u}\|_2 \text{ where } \|\bullet\|_2 \text{ denotes the induced } \ell_2 \text{ matrix norm.}$$

$$(g) \|\mathbf{u}^T \mathbf{v}\|_F \leq \|\mathbf{u}\|_F \|\mathbf{v}\|_F.$$

$$(h) \text{ Let } n = m. \text{ Then, } \|\mathbf{u}\|_2 \leq \|\mathbf{u}\|_F \leq \sqrt{n} \|\mathbf{u}\|_2.$$

$$(i) \text{ For } x \in \mathbb{R}^{n \times 1} \text{ we have } \|\mathbf{u}x\|_2 \leq \|\mathbf{u}\|_F \|x\|_2.$$

Proof is **left as an exercise**.

In-class assignment 1: Let $\mathbf{u} \in \mathbb{R}^{m \times n}$. Let $\bar{\mathbf{u}} \triangleq \frac{1}{m} \mathbf{1}^T \mathbf{u}$ denotes the average of \mathbf{u} across its rows.

Show that

- Let $u \in \mathbb{R}^{1 \times n}$. Then, $\|\mathbf{1}u\|_F^2 = m\|u\|_2^2$
- $\langle \mathbf{u}, \mathbf{1}\bar{\mathbf{u}} \rangle = m\|\bar{\mathbf{u}}\|_2^2$.

$$\|\mathbf{1}u\|_F^2 = \langle \mathbf{1}u, \mathbf{1}u \rangle = \text{Trace}(u^T \mathbf{1}^T \mathbf{1}u) = m \text{Trace}(u^T u) = m \text{Trace}(\|u\|_2^2) = m\|u\|_2^2$$

$$\begin{aligned} \langle \mathbf{u}, \mathbf{1}\bar{\mathbf{u}} \rangle &= \langle \mathbf{1}\bar{\mathbf{u}}, \mathbf{u} \rangle = \langle \mathbf{1} \frac{1}{m} \mathbf{1}^T \mathbf{u}, \mathbf{u} \rangle = \frac{1}{m} \langle \mathbf{1} \mathbf{1}^T \mathbf{u}, \mathbf{u} \rangle = \frac{1}{m} \text{Trace}(\mathbf{u}^T \mathbf{1} \mathbf{1}^T \mathbf{u}) = \frac{1}{m} \text{Trace}((\mathbf{1}^T \mathbf{u})^T \mathbf{1}^T \mathbf{u}) = \frac{1}{m} \langle \mathbf{1}^T \mathbf{u}, \mathbf{1}^T \mathbf{u} \rangle = \frac{1}{m} \text{Trace}((\mathbf{1}^T \mathbf{u})^T \mathbf{1}^T \mathbf{u}) = m \langle \bar{\mathbf{u}}, \bar{\mathbf{u}} \rangle \\ &= m\|\bar{\mathbf{u}}\|_2^2 \end{aligned}$$

Compact representation:

$$\mathbf{x}_{k+1} := \mathbf{W}(\mathbf{x}_k - \gamma_k \mathbf{y}_k)$$

$$\mathbf{y}_{k+1} := \mathbf{W} \mathbf{y}_k + \mathbf{G}(\mathbf{x}_{k+1}, \xi_{k+1}) - \mathbf{G}(\mathbf{x}_k, \xi_k)$$

- The step-size is a scalar and is shared among the agents. It could be diminishing.

Assumption 1:

- The matrix $\mathbf{W} \in \mathbb{R}^{m \times m}$ is doubly stochastic, i.e., $\mathbf{1}^T \mathbf{W} = \mathbf{1}^T$ and $\mathbf{W} \mathbf{1} = \mathbf{1}$.
- The matrix \mathbf{W} is nonnegative.
- For all i , we have $W_{ii} > 0$.

Assumption 2: For all $i \in [m]$, function f_i is μ -strongly convex and L -smooth.

Assumption 3: For all $i \in [m]$ and all $x \in \mathbb{R}^n$, random vectors $\xi_i \in \mathbb{R}^d$ are independent from each other, and

$$\begin{aligned}\mathbb{E} [\nabla f_i(x, \xi_i) | x] &= \nabla f_i(x), \\ \mathbb{E} [\|\nabla f_i(x, \xi_i) - \nabla f_i(x)\|^2 | x] &\leq \nu^2 \quad \text{for some } \nu > 0.\end{aligned}$$

Also, for each agent, the $\{\xi_{i,0}, \xi_{i,1}, \xi_{i,2}, \dots\}$ are an i.i.d. from the random variable ξ_i .

Choosing the weights

To choose the weights and create the matrix \mathbf{W} , if all nodes have the same number of neighbors, then the following formula would satisfy doubly stochasticity assumption:

$$W_{i,j} = \begin{cases} \frac{1-\beta}{|\mathcal{N}_{\mathbf{W}}(i)|}, & \text{if } j \in \mathcal{N}_{\mathbf{W}}(i) \\ \beta, & \text{if } j = i \\ 0, & \text{otherwise} \end{cases}$$

Where $0 \leq \beta \leq 1$ and $|\mathcal{N}_{\mathbf{W}}(i)|$ denotes the number of nodes that are neighbors with node i .

In-class assignment 2:

- Use the above weight rule to obtain \mathbf{W} for an undirected network with 4 agents that communicate over a ring graph. Use $\beta := 0.6$

$$\mathbf{W} = \begin{bmatrix} 0.6 & 0.2 & 0 & 0.2 \\ 0.2 & 0.6 & 0.2 & 0 \\ 0 & 0.2 & 0.6 & 0.2 \\ 0.2 & 0 & 0.2 & 0.6 \end{bmatrix}.$$

- Show that the above-mentioned weight rule satisfies Assumption 1.

Lemma 1: Consider Algorithm 1. Let Assumptions 1-4 hold. Let x^* denote the unique optimal solution of the problem (P). Then, the following hold for all $k \geq 0$:

- (a) $\bar{y}_k = G(\mathbf{x}_k, \xi_k)$. This implies that \bar{y}_k tracks the average of stochastic gradients of the agents at iteration k .
- (b) $\mathbb{E} [\bar{y}_k | \mathcal{F}_k] = G(\mathbf{x}_k)$. This implies that \bar{y}_k is an unbiased estimator of the average of the true gradients of the agents at iteration k .
- (c) $\mathbb{E} [\|\bar{y}_k - G(\mathbf{x}_k)\|^2 | \mathcal{F}_k] \leq \frac{\nu^2}{m}$. This implies that variance of \bar{y}_k is bounded by $\frac{\nu^2}{m}$.
- (d) For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{m \times n}$, with $u_i, v_i \in \mathbb{R}^n$ denoting the i^{th} row of \mathbf{u}, \mathbf{v} , respectively, we have:

$$\|G(\mathbf{u}) - G(\mathbf{v})\| \leq \frac{L}{\sqrt{m}} \|\mathbf{u} - \mathbf{v}\|.$$

$$(e) \|G(\mathbf{x}_k) - \bar{G}(\bar{x}_k)\| \leq \frac{L}{\sqrt{m}} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|.$$

$$(f) \|\bar{G}(\bar{x}_k)\| \leq L \|\bar{x}_k - x^*\|.$$

Proof:

- (a) We use induction to show this statement. For $k = 0$, we have:

$$\bar{y}_0 = \frac{1}{m} \mathbf{1}^T \mathbf{y}_0 = \frac{1}{m} \sum_{i=1}^m y_{i,0} = \frac{1}{m} \sum_{i=1}^m \nabla f_i(x_{i,0}, \xi_{i,0}) = G(\mathbf{x}_0, \xi_0).$$

This implies that the hypothesis statement holds for $k = 0$ due to the initialization of the algorithm. Let us assume that $\bar{y}_k = G(\mathbf{x}_k, \xi_k)$ holds for some $k \geq 0$. We show that it holds true for $k + 1$ as well.

Multiplying both sides of $\mathbf{y}_{k+1} := \mathbf{W}\mathbf{y}_k + \mathbf{G}(\mathbf{x}_{k+1}, \xi_{k+1}) - \mathbf{G}(\mathbf{x}_k, \xi_k)$ by the averaging operator $\frac{1}{m} \mathbf{1}^T$, we have:

$$\begin{aligned}\frac{1}{m} \mathbf{1}^T \mathbf{y}_{k+1} &= \frac{1}{m} \mathbf{1}^T \mathbf{W} \mathbf{y}_k + \frac{1}{m} \mathbf{1}^T \mathbf{G}(\mathbf{x}_{k+1}, \xi_{k+1}) - \frac{1}{m} \mathbf{1}^T \mathbf{G}(\mathbf{x}_k, \xi_k) \\ \Rightarrow \bar{y}_{k+1} &= \frac{1}{m} \mathbf{1}^T \mathbf{y}_k + G(\mathbf{x}_{k+1}, \xi_{k+1}) - G(\mathbf{x}_k, \xi_k) \\ \Rightarrow \bar{y}_{k+1} &= \bar{y}_k + G(\mathbf{x}_{k+1}, \xi_{k+1}) - G(\mathbf{x}_k, \xi_k) \\ \Rightarrow \bar{y}_{k+1} &= G(\mathbf{x}_k, \xi_k) + G(\mathbf{x}_{k+1}, \xi_{k+1}) - G(\mathbf{x}_k, \xi_k) \\ \Rightarrow \bar{y}_{k+1} &= G(\mathbf{x}_{k+1}, \xi_{k+1}).\end{aligned}$$

Hence, the proof of (a) is completed.

(b) Taking conditional expectation on the both sides of the equation in part (a) and invoking the definition of $G(\mathbf{x})$, we obtain:

$$\mathbb{E} [\bar{\mathbf{y}}_k \mid \mathcal{F}_k] = \mathbb{E} [G(\mathbf{x}_k, \boldsymbol{\xi}_k) \mid \mathcal{F}_k] = G(\mathbf{x}_k).$$

(c)

$$\begin{aligned} \mathbb{E} [\|\bar{\mathbf{y}}_k - G(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k] &= \mathbb{E} [\|G(\mathbf{x}_k, \boldsymbol{\xi}_k) - G(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k] \\ &= \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i=1}^m \nabla f_i(x_i, \xi_i) - \frac{1}{m} \sum_{i=1}^m \nabla f_i(x_i) \right\|^2 \mid \mathcal{F}_k \right] \\ &= \frac{1}{m^2} \mathbb{E} \left[\left\| \sum_{i=1}^m (\nabla f_i(x_{i,k}, \xi_{i,k}) - \nabla f_i(x_{i,k})) \right\|^2 \mid \mathcal{F}_k \right] \\ &= \frac{1}{m^2} \mathbb{E} \left[\sum_{i=1}^m \|\nabla f_i(x_{i,k}, \xi_{i,k}) - \nabla f_i(x_{i,k})\|^2 + \sum_{i \neq j} (\nabla f_i(x_{i,k}, \xi_{i,k}) - \nabla f_i(x_{i,k}))^T (\nabla f_j(x_{j,k}, \xi_{j,k}) - \nabla f_j(x_{j,k})) \mid \mathcal{F}_k \right] \\ &= \frac{1}{m^2} \sum_{i=1}^m \mathbb{E} [\|\nabla f_i(x_{i,k}, \xi_{i,k}) - \nabla f_i(x_{i,k})\|^2 \mid \mathcal{F}_k] + \frac{1}{m^2} \sum_{i \neq j} \mathbb{E} [(\nabla f_i(x_{i,k}, \xi_{i,k}) - \nabla f_i(x_{i,k}))^T (\nabla f_j(x_{j,k}, \xi_{j,k}) - \nabla f_j(x_{j,k})) \mid \mathcal{F}_k] \\ &\leq \frac{1}{m^2} \sum_{i=1}^m v^2 + \frac{1}{m^2} \sum_{i \neq j} \mathbb{E} [(\nabla f_i(x_{i,k}, \xi_{i,k}) - \nabla f_i(x_{i,k})) \mid \mathcal{F}_k]^T \mathbb{E} [(\nabla f_j(x_{j,k}, \xi_{j,k}) - \nabla f_j(x_{j,k})) \mid \mathcal{F}_k] \\ &= \frac{v^2}{m}. \end{aligned}$$

Proofs of (d), (e), and (f) are **left as an exercise**.

Lemma 2: Let Assumption 2 hold. For any $\alpha \leq \frac{2}{\mu+L}$, we have:

$$\|\bar{\mathbf{x}}_k - \alpha \bar{G}(\bar{\mathbf{x}}_k) - \mathbf{x}^*\| \leq (1 - \mu\alpha) \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|.$$

Consider the first update rule:

$$\mathbf{x}_{k+1} := \mathbf{W} (\mathbf{x}_k - \gamma_k \mathbf{y}_k).$$

Multiplying both sides by the averaging operator $\frac{1}{m} \mathbf{1}^T$ and noting that $\mathbf{1}^T \mathbf{W} = \mathbf{1}^T$, we obtain:

$$\bar{\mathbf{x}}_{k+1} = \bar{\mathbf{x}}_k - \gamma_k \bar{\mathbf{y}}_k.$$

Let us consider the error metric $\mathbb{E} [\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}^*\|^2]$. Using Lemma 1(b) and (c), we can write:

$$\begin{aligned} \mathbb{E} [\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}^*\|^2 \mid \mathcal{F}_k] &= \mathbb{E} [\|\bar{\mathbf{x}}_k - \gamma_k \bar{\mathbf{y}}_k - \mathbf{x}^*\|^2 \mid \mathcal{F}_k] \\ &= \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 - 2\gamma_k (\bar{\mathbf{x}}_k - \mathbf{x}^*)^T \mathbb{E} [\bar{\mathbf{y}}_k \mid \mathcal{F}_k] + \gamma_k^2 \mathbb{E} [\|\bar{\mathbf{y}}_k\|^2 \mid \mathcal{F}_k] \\ &= \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 - 2\gamma_k (\bar{\mathbf{x}}_k - \mathbf{x}^*)^T G(\mathbf{x}_k) + \gamma_k^2 \mathbb{E} [\|\bar{\mathbf{y}}_k - G(\mathbf{x}_k) + G(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k] \\ &= \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 - 2\gamma_k (\bar{\mathbf{x}}_k - \mathbf{x}^*)^T G(\mathbf{x}_k) + \gamma_k^2 \mathbb{E} [\|\bar{\mathbf{y}}_k - G(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k] + \gamma_k^2 \mathbb{E} [\|G(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k] + 2\gamma_k^2 G(\mathbf{x}_k)^T \mathbb{E} [\bar{\mathbf{y}}_k - G(\mathbf{x}_k) \mid \mathcal{F}_k] \\ &= \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 - 2\gamma_k (\bar{\mathbf{x}}_k - \mathbf{x}^*)^T G(\mathbf{x}_k) + \gamma_k^2 \mathbb{E} [\|\bar{\mathbf{y}}_k - G(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k] + \gamma_k^2 \|G(\mathbf{x}_k)\|^2 \\ &\leq \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 - 2\gamma_k (\bar{\mathbf{x}}_k - \mathbf{x}^*)^T G(\mathbf{x}_k) + \gamma_k^2 \frac{v^2}{m} + \gamma_k^2 \|G(\mathbf{x}_k)\|^2. \end{aligned}$$

Adding and subtracting $\bar{G}(\bar{\mathbf{x}}_k)$, we obtain:

$$\begin{aligned} \mathbb{E} [\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}^*\|^2 \mid \mathcal{F}_k] &\leq \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 - 2\gamma_k (\bar{\mathbf{x}}_k - \mathbf{x}^*)^T (G(\mathbf{x}_k) - \bar{G}(\bar{\mathbf{x}}_k)) - 2\gamma_k (\bar{\mathbf{x}}_k - \mathbf{x}^*)^T \bar{G}(\bar{\mathbf{x}}_k) + \gamma_k^2 \frac{v^2}{m} \\ &\quad + \gamma_k^2 \|G(\mathbf{x}_k) - \bar{G}(\bar{\mathbf{x}}_k)\|^2 + \gamma_k^2 \|\bar{G}(\bar{\mathbf{x}}_k)\|^2 + 2\gamma_k^2 (G(\mathbf{x}_k) - \bar{G}(\bar{\mathbf{x}}_k))^T \bar{G}(\bar{\mathbf{x}}_k) \\ &\leq \|\bar{\mathbf{x}}_k - \mathbf{x}^* - \gamma_k \bar{G}(\bar{\mathbf{x}}_k)\|^2 - 2\gamma_k (\bar{\mathbf{x}}_k - \mathbf{x}^*)^T (G(\mathbf{x}_k) - \bar{G}(\bar{\mathbf{x}}_k)) + \gamma_k^2 \frac{v^2}{m} \\ &\quad + \gamma_k^2 \|G(\mathbf{x}_k) - \bar{G}(\bar{\mathbf{x}}_k)\|^2 + 2\gamma_k^2 (G(\mathbf{x}_k) - \bar{G}(\bar{\mathbf{x}}_k))^T \bar{G}(\bar{\mathbf{x}}_k). \end{aligned}$$

Invoking Lemma 2, we obtain:

$$\begin{aligned} \mathbb{E} [\|\bar{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k] &\leq (1 - \mu\gamma_k)^2 \|\bar{x}_k - x^*\|^2 - 2\gamma_k (\bar{x}_k - \gamma_k \bar{G}(\bar{x}_k) - x^*)^T (G(\mathbf{x}_k) - \bar{G}(\bar{x}_k)) \\ &\quad + \gamma_k^2 \frac{v^2}{m} + \gamma_k^2 \|G(\mathbf{x}_k) - \bar{G}(\bar{x}_k)\|^2. \end{aligned}$$

Invoking the Cauchy-Schwarz inequality and Lemma 2 again, we obtain:

$$\begin{aligned} \mathbb{E} [\|\bar{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k] &\leq (1 - \mu\gamma_k)^2 \|\bar{x}_k - x^*\|^2 + 2\gamma_k (1 - \mu\gamma_k) \|\bar{x}_k - x^*\| \|G(\mathbf{x}_k) - \bar{G}(\bar{x}_k)\| \\ &\quad + \gamma_k^2 \frac{v^2}{m} + \gamma_k^2 \|G(\mathbf{x}_k) - \bar{G}(\bar{x}_k)\|^2. \end{aligned}$$

Using Lemma 1(e), we obtain:

$$\begin{aligned} \mathbb{E} [\|\bar{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k] &\leq (1 - \mu\gamma_k)^2 \|\bar{x}_k - x^*\|^2 + \frac{2\gamma_k L(1 - \mu\gamma_k)}{\sqrt{m}} \|\bar{x}_k - x^*\| \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\| \\ &\quad + \gamma_k^2 \frac{v^2}{m} + \frac{\gamma_k^2 L^2}{m} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2. \end{aligned}$$

Note that we have:

$$\begin{aligned} \frac{2\gamma_k L(1 - \mu\gamma_k)}{\sqrt{m}} \|\bar{x}_k - x^*\| \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\| &= 2\gamma_k \left(\sqrt{\mu}(1 - \mu\gamma_k) \|\bar{x}_k - x^*\| \right) \left(\frac{L}{\sqrt{\mu m}} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\| \right) \\ &\leq \gamma_k \left(\mu(1 - \mu\gamma_k)^2 \|\bar{x}_k - x^*\|^2 + \frac{L^2}{\mu m} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 \right). \end{aligned}$$

From the preceding two relations, we obtain:

$$\mathbb{E} [\|\bar{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq (1 - \mu\gamma_k)^2 (1 + \mu\gamma_k) \|\bar{x}_k - x^*\|^2 + \frac{\gamma_k L^2}{\mu m} (1 + \mu\gamma_k) \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \gamma_k^2 \frac{v^2}{m}.$$

Therefore, we obtain the **first recursion**:

$$\mathbb{E} [\|\bar{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq (1 - \mu\gamma_k) \|\bar{x}_k - x^*\|^2 + \frac{\gamma_k L^2}{\mu m} (1 + \mu\gamma_k) \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \gamma_k^2 \frac{v^2}{m}.$$

Assumption 4: Let the graph \mathcal{G} corresponding to the communication network be undirected and connected.

Lemma 3: Let Assumption 1 hold. Let ρ_W denote the spectral norm of the matrix $\mathbf{W} - \frac{1}{m} \mathbf{1}\mathbf{1}^T$.

Then, $\rho_W < 1$ and

$$\|\mathbf{W}\mathbf{u} - \mathbf{1}\bar{u}\| \leq \rho_W \|\mathbf{u} - \mathbf{1}\bar{u}\| \quad \text{for all } \mathbf{u} \in \mathbb{R}^{m \times n},$$

where $\bar{u} \triangleq \frac{1}{m} \mathbf{1}^T \mathbf{u}$.

```
In [27]: 1 m= 3
          2 n= 5
          3 u = np.random.randint(5, size=(m, n))
          4 print("u = \n",u,"\n")
          5 ones = np.ones([m,1])
          6
          7 bar_u = (1/m)*ones.T@u
          8 print("bar_u = \n",bar_u,"\n")
          9
          10 ones_bar_u = ones@bar_u
          11 print("ones_bar_u = \n",ones@bar_u)
```

```
u =
[[0 1 1 4 4]
 [0 1 3 1 2]
 [3 0 1 1 0]]

bar_u =
[[1.          0.66666667 1.66666667 2.          2.          ]

ones_bar_u =
[[1.          0.66666667 1.66666667 2.          2.          ]
 [1.          0.66666667 1.66666667 2.          2.          ]
 [1.          0.66666667 1.66666667 2.          2.          ]]
```

The second recursion

Here, we find a recursive relation for the term $\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2$. Invoking Lemma 2(b), we have

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 &= \|\mathbf{W}\mathbf{x}_k - \gamma_k \mathbf{W}\mathbf{y}_k - \mathbf{1}(\bar{x}_k - \gamma_k \bar{y}_k)\|^2 \\
&= \|\mathbf{W}\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 - 2\gamma_k \langle \mathbf{W}\mathbf{x}_k - \mathbf{1}\bar{x}_k, \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k \rangle + \gamma_k^2 \|\mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2.
\end{aligned}$$

Invoking Lemma 3 and Lemma 2(c), we obtain

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 &= \rho_W^2 \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + 2\gamma_k \|\mathbf{W}\mathbf{x}_k - \mathbf{1}\bar{x}_k\| \|\mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k\| + \rho_W^2 \gamma_k^2 \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\
&\leq \rho_W^2 \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + 2\rho_W^2 \gamma_k \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\| \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\| + \rho_W^2 \gamma_k^2 \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\
&\leq \rho_W^2 \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \rho_W^2 \gamma_k \left(\frac{1 - \rho_W^2}{2\gamma_k \rho_W^2} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{2\gamma_k \rho_W^2}{1 - \rho_W^2} \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \right) + \rho_W^2 \gamma_k^2 \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\
&= \frac{1 + \rho_W^2}{2} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{\gamma_k^2 (1 + \rho_W^2) \rho_W^2}{1 - \rho_W^2} \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2.
\end{aligned}$$

So, taking expectations from both sides, the second recursion is as follows:

$$\mathbb{E} [\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2] \leq \frac{1 + \rho_W^2}{2} \mathbb{E} [\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2] + \frac{\gamma_k^2 (1 + \rho_W^2) \rho_W^2}{1 - \rho_W^2} \mathbb{E} [\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2]$$

Next we obtain the third recursive relation. For the ease of presentation, we use the compact notation

$$\mathbf{G}_k \triangleq \mathbf{G}(\mathbf{x}_k),$$

$$\tilde{\mathbf{G}}_k \triangleq \mathbf{G}(\mathbf{x}_k, \xi_k),$$

$$\nabla_{i,k}^f \triangleq \nabla f_i(x_{i,k}),$$

$$\tilde{\nabla}_{i,k}^f \triangleq \nabla f_i(x_{i,k}, \xi_{i,k}).$$

From the update rules of the algorithm, we have

$$\begin{aligned}
\|\mathbf{y}_{k+1} - \mathbf{1}\bar{y}_{k+1}\|^2 & \leq \|\mathbf{W}\mathbf{y}_k + \tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k - \mathbf{1}\bar{y}_k + \mathbf{1}\bar{y}_k - \mathbf{1}\bar{y}_{k+1}\|^2 \\
& = \|\mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 + \|\tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k\|^2 \\
& \quad + 2\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, \tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k \rangle \\
& \quad + 2\langle \mathbf{y}_{k+1} - \mathbf{1}\bar{y}_k, \mathbf{1}(\bar{y}_k - \bar{y}_{k+1}) \rangle + m\|\bar{y}_k - \bar{y}_{k+1}\|^2 \\
& = \rho_W^2 \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 + \|\tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k\|^2 \\
& \quad + 2\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, \tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k \rangle - m\|\bar{y}_k - \bar{y}_{k+1}\|^2 \\
& \leq \rho_W^2 \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 + \|\tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k\|^2 \\
& \quad + 2\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, \tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k \rangle.
\end{aligned} \tag{Eq. 1}$$

In the following, we present a few intermediary results that will be used to derive the third recursive inequality.

Claim 1: The following holds

$$\begin{aligned}
\mathbb{E} [\|\tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k\|^2 \mid \mathcal{F}_k] &\leq \mathbb{E} [\|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2 \mid \mathcal{F}_k] \\
&\quad + 2\mathbb{E} [\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}_k + \mathbf{G}_k \rangle \mid \mathcal{F}_k] + 2mv^2.
\end{aligned}$$

Proof of Claim 1: We can write

$$\begin{aligned}
\mathbb{E} [\|\tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k\|^2 \mid \mathcal{F}_k] &= \mathbb{E} [\|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2 \mid \mathcal{F}_k] \\
&\quad + 2\mathbb{E} [\langle \mathbf{G}_{k+1}, \tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k - \mathbf{G}_{k+1} + \mathbf{G}_k \rangle \mid \mathcal{F}_k] \\
&\quad - 2\mathbb{E} [\langle \mathbf{G}_k, \tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k - \mathbf{G}_{k+1} + \mathbf{G}_k \rangle \mid \mathcal{F}_k] \\
&\quad + \mathbb{E} [\|\tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k - \mathbf{G}_{k+1} + \mathbf{G}_k\|^2 \mid \mathcal{F}_k].
\end{aligned}$$

Note that since \mathbf{x}_{k+1} is characterized in terms of ξ_k , we have

$$\begin{aligned}
&\mathbb{E} [\tilde{\mathbf{G}}_{k+1} - \mathbf{G}_{k+1} \mid \mathcal{F}_k] \\
&= \mathbb{E}_{\xi_k} [\mathbb{E}_{\xi_{k+1}} [\tilde{\mathbf{G}}_{k+1} - \mathbf{G}_{k+1} \mid \mathcal{F}_{k+1}]] = \mathbb{E}_{\xi_k} [0] = 0.
\end{aligned}$$

Similarly, $\mathbb{E} [\tilde{\mathbf{G}}_k - \mathbf{G}_k \mid \mathcal{F}_k] = 0$. Thus, we obtain

$$\mathbb{E} [\langle \mathbf{G}_k, \tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k - \mathbf{G}_{k+1} + \mathbf{G}_k \rangle \mid \mathcal{F}_k] = 0.$$

We can also write

$$\begin{aligned}
& \mathbb{E} \left[\langle \mathbf{G}_{k+1}, \tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k - \mathbf{G}_{k+1} + \mathbf{G}_k \rangle \mid \mathcal{F}_k \right] \\
&= \mathbb{E}_{\xi_k} \left[\mathbb{E}_{\xi_{k+1}} \left[\langle \mathbf{G}_{k+1}, \tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k - \mathbf{G}_{k+1} + \mathbf{G}_k \rangle \mid \mathcal{F}_{k+1} \right] \right] \\
&= \mathbb{E}_{\xi_k} \left[\langle \mathbf{G}_{k+1}, \mathbf{G}_{k+1} - \tilde{\mathbf{G}}_k - \mathbf{G}_{k+1} + \mathbf{G}_k \rangle \mid \mathcal{F}_k \right] \\
&= \mathbb{E} \left[\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}_k + \mathbf{G}_k \rangle \mid \mathcal{F}_k \right]
\end{aligned}$$

From the preceding relations, we have

$$\begin{aligned}
& \mathbb{E} \left[\|\tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k\|^2 \mid \mathcal{F}_k \right] \leq \mathbb{E} \left[\|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2 \mid \mathcal{F}_k \right] \\
&+ 2\mathbb{E} \left[\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}_k + \mathbf{G}_k \rangle \mid \mathcal{F}_k \right] \\
&+ \mathbb{E} \left[\|\tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k - \mathbf{G}_{k+1} + \mathbf{G}_k\|^2 \mid \mathcal{F}_k \right] \\
&\leq \mathbb{E} \left[\|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2 \mid \mathcal{F}_k \right] + 2\mathbb{E} \left[\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}_k + \mathbf{G}_k \rangle \mid \mathcal{F}_k \right] \\
&+ \mathbb{E} \left[\|\tilde{\mathbf{G}}_{k+1} - \mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k \right] + \mathbb{E} \left[\|\tilde{\mathbf{G}}_k - \mathbf{G}_k\|^2 \mid \mathcal{F}_k \right] + 2\mathbb{E} \left[\langle \tilde{\mathbf{G}}_{k+1} - \mathbf{G}_{k+1}, \tilde{\mathbf{G}}_k - \mathbf{G}_k \rangle \mid \mathcal{F}_k \right].
\end{aligned}$$

It suffices to show that

$$\mathbb{E} \left[\|\tilde{\mathbf{G}}_{k+1} - \mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k \right] \leq mv^2,$$

Claim 2: The following holds

$$\mathbb{E} \left[\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}_k + \mathbf{G}_k \rangle \mid \mathcal{F}_k \right] \leq m\gamma_k L v^2.$$

Proof of Claim 2: Let us define for all $i \in [m]$

$$\hat{\mathbf{x}}_{i,k+1} \triangleq \mathbf{x}_{i,k+1} + \gamma_k W_{ii} \left(\nabla f_i(\mathbf{x}_{i,k}, \xi_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}) \right).$$

We can write

$$\begin{aligned}
& \mathbb{E} \left[\langle \nabla_{i,k+1}^f, -\tilde{\nabla}_{i,k}^f - \nabla_{i,k}^f \rangle \mid \mathcal{F}_k \right] \\
&= \mathbb{E} \left[\langle \nabla_{i,k+1}^f - \nabla f_i(\hat{\mathbf{x}}_{i,k+1}), -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k \right] \\
&+ \mathbb{E} \left[\langle \nabla f_i(\hat{\mathbf{x}}_{i,k+1}), -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k \right].
\end{aligned}$$

It can be shown from the update rules of the algorithm that $\hat{\mathbf{x}}_{i,k+1}$ is independent of $\xi_{i,k}$ (proof is **left as an exercise**). Thus, we can write

$$\mathbb{E} \left[\langle \nabla f_i(\hat{\mathbf{x}}_{i,k+1}), -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k \right] = 0.$$

Also, from the Lipschitzian property and the definition of $\hat{\mathbf{x}}_{i,k+1}$ we have

$$\|\nabla_{i,k+1}^f - \nabla f_i(\hat{\mathbf{x}}_{i,k+1})\| \leq L\|\mathbf{x}_{i,k+1} - \hat{\mathbf{x}}_{i,k+1}\| \leq \gamma_k L \|\tilde{\nabla}_{i,k}^f - \nabla_{i,k}^f\|.$$

From the preceding relations, we have

$$\begin{aligned}
& \mathbb{E} \left[\langle \nabla_{i,k+1}^f, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k \right] \\
&\leq \mathbb{E} \left[\|\nabla_{i,k+1}^f - \nabla f_i(\hat{\mathbf{x}}_{i,k+1})\| \|\tilde{\nabla}_{i,k}^f - \nabla_{i,k}^f\| \mid \mathcal{F}_k \right] \\
&\leq \gamma_k L \mathbb{E} \left[\|\tilde{\nabla}_{i,k}^f - \nabla_{i,k}^f\|^2 \mid \mathcal{F}_k \right] \\
&\leq \gamma_k L v^2.
\end{aligned}$$

Summing from the preceding relation over i , we obtain Claim 2.

Claim 3: The following holds

$$\begin{aligned}
& \|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2 \\
&\leq L^2 (3L^2 \gamma_k^2 + 2\|\mathbf{W} - \mathbf{I}\|^2) \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 \\
&+ 2L^2 \rho_W^2 \gamma_k^2 \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\
&+ 3mL^4 \gamma_k^2 \|\bar{x}_k - x^*\|^2 \\
&+ 3L^2 v^2 \gamma_k^2.
\end{aligned}$$

Proof of Claim 3: From the Lipschitzian property of the local objective functions we have (proof is **left as an exercise**)

$$\|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2 \leq L^2 \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.$$

We also have

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 &= \|\mathbf{W}\mathbf{x}_k - \gamma_k \mathbf{W}\mathbf{y}_k - \mathbf{x}_k\|^2 \\
&= \|(\mathbf{W} - \mathbf{I})(\mathbf{x}_k - \mathbf{1}\bar{x}_k) - \gamma_k \mathbf{W}\mathbf{y}_k - \mathbf{x}_k\|^2 \\
&\leq \|\mathbf{W} - \mathbf{I}\|^2 \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \gamma_k^2 \|\mathbf{W}\mathbf{y}_k\|^2 \\
&\quad - 2\gamma_k \langle (\mathbf{W} - \mathbf{I})(\mathbf{x}_k - \mathbf{1}\bar{x}_k), \mathbf{W}\mathbf{y}_k \rangle \\
&= \|\mathbf{W} - \mathbf{I}\|^2 \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \gamma_k^2 \|\mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 + m\gamma_k^2 \|\bar{y}_k\|^2 \\
&\quad - 2\gamma_k \langle (\mathbf{W} - \mathbf{I})(\mathbf{x}_k - \mathbf{1}\bar{x}_k), \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k \rangle \\
&\leq \|\mathbf{W} - \mathbf{I}\|^2 \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \rho_W^2 \gamma_k^2 \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 + m\gamma_k^2 \|\bar{y}_k\|^2 \\
&\quad + 2\rho_W \gamma_k \|\mathbf{W} - \mathbf{I}\| \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\| \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\| \\
&\leq 2\|\mathbf{W} - \mathbf{I}\|^2 \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + 2\rho_W^2 \gamma_k^2 \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\
&\quad + m\gamma_k^2 \|\bar{y}_k\|^2.
\end{aligned}$$

$$\|x + y + z\|^2 \leq 3\|x\|^2 + 3\|y\|^2 + 3\|z\|^2$$

Claim 4: The following holds

$$\begin{aligned}
&\mathbb{E} [\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, \tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k \rangle \mid \mathcal{F}_k] \\
&= \mathbb{E} [\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, \mathbf{G}_{k+1} - \mathbf{G}_k \rangle \mid \mathcal{F}_k] \\
&\quad + \mathbb{E} [\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, -\tilde{\mathbf{G}}_k + \mathbf{G}_k \rangle \mid \mathcal{F}_k].
\end{aligned}$$

Proof of Claim 4: We have

$$\begin{aligned}
&\mathbb{E} [\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, \tilde{\mathbf{G}}_{k+1} - \mathbf{G}_{k+1} \rangle \mid \mathcal{F}_k] \\
&= \mathbb{E}_{\xi_k} [\mathbb{E}_{\xi_{k+1}} [\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, \tilde{\mathbf{G}}_{k+1} - \mathbf{G}_{k+1} \rangle \mid \mathcal{F}_{k+1}]] = 0.
\end{aligned}$$

The result follows by adding the above expectation to the left-hand side of Claim 4.

Claim 5: The following holds

$$\mathbb{E} [\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, -\tilde{\mathbf{G}}_k + \mathbf{G}_k \rangle \mid \mathcal{F}_k] \leq v^2.$$

Proof of Claim 5: First, note that from the algorithm's update rules, we have for any $i, j \in [m]$

$$\begin{aligned}
&\mathbb{E} [\langle y_{j,k}, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k] \\
&= \mathbb{E} [\langle \sum_{\ell=1}^m W_{j\ell} y_{\ell,k-1} + \tilde{\nabla}_{j,k}^f - \tilde{\nabla}_{j,k-1}^f, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k] \\
&= \mathbb{E} [\langle \tilde{\nabla}_{j,k}^f, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k].
\end{aligned}$$

Multiplying by W_{ij} and summing over $j \in [m]$, we have

$$\begin{aligned}
&\mathbb{E} [\langle \sum_{j=1}^m W_{ij} y_{j,k}, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k] \\
&= \mathbb{E} [\langle W_{ii} \tilde{\nabla}_{i,k}^f, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k] \\
&= W_{ii} \mathbb{E} [\langle \tilde{\nabla}_{i,k}^f - \nabla_{i,k}^f, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k] \leq 0.
\end{aligned}$$

We have

$$\begin{aligned}
&-\mathbb{E} [\langle \bar{y}_k, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k] \\
&= -\mathbb{E} [\langle \frac{1}{m} \sum_{j=1}^m \tilde{\nabla}_{j,k}^f, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k] \\
&= \frac{1}{m} \mathbb{E} [\langle -\tilde{\nabla}_{i,k}^f, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k] \\
&= \frac{1}{m} \mathbb{E} [\langle -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k] \\
&= \frac{1}{m} \mathbb{E} [\|\nabla_{i,k}^f - \tilde{\nabla}_{i,k}^f\|^2 \mid \mathcal{F}_k] \\
&\leq \frac{v^2}{m}.
\end{aligned}$$

Employing the preceding two relations, we have

$$\begin{aligned}
&\mathbb{E} [\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, -\tilde{\mathbf{G}}_k + \mathbf{G}_k \rangle \mid \mathcal{F}_k] \\
&= \sum_{i=1}^m \mathbb{E} [\langle \sum_{j=1}^m W_{ij} y_{j,k} - \bar{y}_k, -\tilde{\nabla}_{i,k}^f + \nabla_{i,k}^f \rangle \mid \mathcal{F}_k] \\
&\leq \sum_{i=1}^m \frac{v^2}{m} = v^2.
\end{aligned}$$

This implies that Claim 5 holds.

Claim 6: The following holds for any $\eta > 0$

$$\begin{aligned}
&2\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, \mathbf{G}_{k+1} - \mathbf{G}_k \rangle \\
&\leq \eta \rho_W^2 \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 + \eta^{-1} \|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2.
\end{aligned}$$

From Claim 4 and Eq. 1, we have

$$\begin{aligned} & \mathbb{E} [\|\mathbf{y}_{k+1} - \mathbf{1}\bar{\mathbf{y}}_{k+1}\|^2 \mid \mathcal{F}_k] \\ & \leq \rho_W^2 \|\mathbf{y}_k - \mathbf{1}\bar{\mathbf{y}}_k\|^2 + \mathbb{E} [\|\tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k\|^2 \mid \mathcal{F}_k] \\ & + 2\mathbb{E} [\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{\mathbf{y}}_k, \mathbf{G}_{k+1} - \mathbf{G}_k \rangle \mid \mathcal{F}_k] \\ & + 2\mathbb{E} [\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{\mathbf{y}}_k, -\tilde{\mathbf{G}}_k + \mathbf{G}_k \rangle \mid \mathcal{F}_k]. \end{aligned}$$

Claim 5 and 6 imply that

$$\begin{aligned} & \mathbb{E} [\|\mathbf{y}_{k+1} - \mathbf{1}\bar{\mathbf{y}}_{k+1}\|^2 \mid \mathcal{F}_k] \\ & \leq (1 + \eta)\rho_W^2 \mathbb{E} [\|\mathbf{y}_k - \mathbf{1}\bar{\mathbf{y}}_k\|^2 \mid \mathcal{F}_k] \\ & + (1 + \eta^{-1})\mathbb{E} [\|\tilde{\mathbf{G}}_{k+1} - \tilde{\mathbf{G}}_k\|^2 \mid \mathcal{F}_k] \\ & + 2v^2. \end{aligned}$$

From Claim 1 we obtain

$$\begin{aligned} & \mathbb{E} [\|\mathbf{y}_{k+1} - \mathbf{1}\bar{\mathbf{y}}_{k+1}\|^2 \mid \mathcal{F}_k] \\ & \leq (1 + \eta)\rho_W^2 \mathbb{E} [\|\mathbf{y}_k - \mathbf{1}\bar{\mathbf{y}}_k\|^2 \mid \mathcal{F}_k] \\ & + (1 + \eta^{-1})\mathbb{E} [\|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2 \mid \mathcal{F}_k] \\ & + 2(1 + \eta^{-1})\mathbb{E} [\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}_k + \mathbf{G}_k \rangle \mid \mathcal{F}_k] \\ & + 2(1 + \eta^{-1})mv^2 \\ & + 2v^2. \end{aligned}$$

Claim 2 implies

$$\begin{aligned} & \mathbb{E} [\|\mathbf{y}_{k+1} - \mathbf{1}\bar{\mathbf{y}}_{k+1}\|^2 \mid \mathcal{F}_k] \\ & \leq (1 + \eta)\rho_W^2 \mathbb{E} [\|\mathbf{y}_k - \mathbf{1}\bar{\mathbf{y}}_k\|^2 \mid \mathcal{F}_k] \\ & + (1 + \eta^{-1})L^2 (3L^2\gamma_k^2 + 2\|\mathbf{W} - \mathbf{I}\|^2) \|\mathbf{x}_k - \mathbf{1}\bar{\mathbf{x}}_k\|^2 \\ & + (1 + \eta^{-1})2L^2\rho_W^2\gamma_k^2 \mathbb{E} [\|\mathbf{y}_k - \mathbf{1}\bar{\mathbf{y}}_k\|^2 \mid \mathcal{F}_k] \\ & + (1 + \eta^{-1})3mL^4\gamma_k^2 \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 \\ & + (1 + \eta^{-1})3L^2v^2\gamma_k^2 \\ & + 2(1 + \eta^{-1})m\gamma_kLv^2 \\ & + 2(1 + \eta^{-1})mv^2 \\ & + 2v^2. \end{aligned}$$

Rearranging the terms and taking expectations from both sides, we obtain

$$\begin{aligned} & \mathbb{E} [\|\mathbf{y}_{k+1} - \mathbf{1}\bar{\mathbf{y}}_{k+1}\|^2] \\ & \leq (1 + \eta + (1 + \eta^{-1})2L^2\gamma_k^2) \rho_W^2 \mathbb{E} [\|\mathbf{y}_k - \mathbf{1}\bar{\mathbf{y}}_k\|^2] \\ & + (1 + \eta^{-1})L^2 (3L^2\gamma_k^2 + 2\|\mathbf{W} - \mathbf{I}\|^2) \mathbb{E} [\|\mathbf{x}_k - \mathbf{1}\bar{\mathbf{x}}_k\|^2] \\ & + (1 + \eta^{-1})3mL^4\gamma_k^2 \mathbb{E} [\|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2] \\ & + ((1 + \eta^{-1}) (3L^2\gamma_k^2 + 2m\gamma_kL + 2m) + 2) v^2. \end{aligned}$$

This implies the third recursive inequality.

To summarize the three recursions are as follows

$$\begin{aligned} & \mathbb{E} [\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}^*\|^2] \leq (1 - \mu\gamma_k) \mathbb{E} [\|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2] + \frac{\gamma_k L^2}{\mu m} (1 + \mu\gamma_k) \mathbb{E} [\|\mathbf{x}_k - \mathbf{1}\bar{\mathbf{x}}_k\|^2] + \gamma_k^2 \frac{v^2}{m}, \\ & \mathbb{E} [\|\mathbf{x}_{k+1} - \mathbf{1}\bar{\mathbf{x}}_{k+1}\|^2] \leq \frac{1 + \rho_W^2}{2} \mathbb{E} [\|\mathbf{x}_k - \mathbf{1}\bar{\mathbf{x}}_k\|^2] + \frac{\gamma_k^2 (1 + \rho_W^2) \rho_W^2}{1 - \rho_W^2} \mathbb{E} [\|\mathbf{y}_k - \mathbf{1}\bar{\mathbf{y}}_k\|^2], \\ & \mathbb{E} [\|\mathbf{y}_{k+1} - \mathbf{1}\bar{\mathbf{y}}_{k+1}\|^2] \leq (1 + \eta + (1 + \eta^{-1})2L^2\gamma_k^2) \rho_W^2 \mathbb{E} [\|\mathbf{y}_k - \mathbf{1}\bar{\mathbf{y}}_k\|^2] \\ & + (1 + \eta^{-1})L^2 (3L^2\gamma_k^2 + 2\|\mathbf{W} - \mathbf{I}\|^2) \mathbb{E} [\|\mathbf{x}_k - \mathbf{1}\bar{\mathbf{x}}_k\|^2] + (1 + \eta^{-1})3mL^4\gamma_k^2 \mathbb{E} [\|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2] \\ & + ((1 + \eta^{-1}) (3L^2\gamma_k^2 + 2m\gamma_kL + 2m) + 2) v^2, \end{aligned}$$

where $\eta > 0$ is an arbitrary scalar.

Choice of η

To guarantee the convergence, it is important to choose η to have that

$$(1 + \eta + (1 + \eta^{-1})2L^2\gamma_k^2) \rho_W^2 < 1.$$

Let us assume

$$(1 + \eta + (1 + \eta^{-1})2L^2\gamma_k^2) \rho_W^2 \leq \frac{1 + \rho_W^2}{2}.$$

Equivalently, we need to have

$$\eta + (1 + \eta^{-1})2L^2\gamma_k^2 \leq \frac{1 - \rho_W^2}{2\rho_W^2}.$$

Let us set $\eta := \frac{1 - \rho_W^2}{4\rho_W^2}$. We must have

$$(1 + \eta^{-1})2L^2\gamma_k^2 \leq \frac{1 - \rho_W^2}{4\rho_W^2}.$$

We must have

$$\gamma_k \leq \frac{1 - \rho_W^2}{2L\rho_W\sqrt{2 + 6\rho_W^2}}.$$

This can hold if we set

$$\gamma_0 = \frac{1 - \rho_W^2}{2L\rho_W\sqrt{2 + 6\rho_W^2}},$$

and

$$\gamma_k \triangleq \frac{\gamma_0}{k + 1}.$$

Convergence Rate of DSGT

It can be shown that there exist some $E_1, E_2, E_3 > 0$ such that for $k \geq 0$

$$\begin{aligned} \mathbb{E} [\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}^*\|^2] &\leq \frac{E_1}{k + 1} \\ \mathbb{E} [\|\mathbf{x}_{k+1} - \mathbf{1}\bar{\mathbf{x}}_{k+1}\|^2] &\leq \frac{E_2}{(k + 1)^2}, \\ \mathbb{E} [\|\mathbf{y}_{k+1} - \mathbf{1}\bar{\mathbf{y}}_{k+1}\|^2] &\leq E_3. \end{aligned} \tag{Eq. 2}$$

Proof sketch

We use induction on k . Let us assume Eq. 2 holds for k . We obtain

$$\mathbb{E} [\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}^*\|^2] \leq (1 - \mu\gamma_k) \frac{E_1}{k} + \frac{\gamma_k L^2}{\mu m} (1 + \mu\gamma_k) \frac{E_2}{k^2} + \gamma_k^2 \frac{v^2}{m}.$$

Thus, it would suffice to show

$$(1 - \mu\gamma_k) \frac{E_1}{k} + \frac{\gamma_k L^2}{\mu m} (1 + \mu\gamma_k) \frac{E_2}{k^2} + \gamma_k^2 \frac{v^2}{m} \leq \frac{E_1}{k + 1}.$$

Equivalently,

$$\frac{\gamma_k L^2}{\mu m} (1 + \mu\gamma_k) \frac{E_2}{k^2} + \gamma_k^2 \frac{v^2}{m} \leq \left(\frac{1}{k + 1} - 1 + \frac{\mu\gamma_0}{k} \right) E_1.$$