

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет комп'ютерних наук та кібернетики

Кафедра обчислювальної математики

Кваліфікаційна робота
на здобуття ступеня магістра
за спеціальністю 113 Прикладна математика
на тему:

**ПРИСКОРЕНІ ГРАДІЄНТНІ МЕТОДИ.
НЕПЕРЕРВНІ ВАРІАНТИ ПРИСКОРЕНИХ МЕТОДІВ. ФУНКЦІЇ
ЛЯПУНОВА.**

Керівник дипломної роботи
професор
Клюшин Д. А.

« ____ » _____ 2023р.

Виконав студент 2го курсу
Ковальчук В. О.

« ____ » _____ 2023р.

Роботу розглянуто й допущено до захисту на засідання кафедри
обчислювальної математикм від "5" травня 2023р., протокол засідання №7
Завідувач кафедри обчислювальної математики

проф. Сергій Ляшко



Київ 2023

Зміст

1 Вступ	2
2 Мета роботи	3
3 Розглянуті методи	3
3.1 Градієнтний метод	3
3.2 Градієнтний метод для μ -сильно випуклих функцій	6
4 Лінійний каплінг	9
4.1 Дзеркальний спуск	9
4.2 Алгоритм лінійного каплінгу	11
5 Приклади	14
5.1 Розглянка Функція: $f = \frac{x^2 * n}{2}$	14
5.2 Розглянка Функція: $f(x, y) = \frac{1}{2} \frac{x^2 + ny^2}{n+1}$	17
6 Неперервний випадок	20
6.1 Додаткові теоретичні результати	22
6.2 Метою важкої кульки. Загальний вигляд	23
6.3 Метод важкої кульки у гамільтоновій формі.	26
7 Приклад з логістичною регресією	28
7.1 Приклад для двовимірного розділення точок	28
7.2 Приклад для тривимірного розділення точок	33
8 Лінійна регресія	35
9 Висновок	38

1 Вступ

У цьому дослідженні ми ознайомимось із кількома методами, що застосовуються до задачі опуклого програмування. Наша основна увага буде приділена оптимізації, яка полягає в пошуку мінімуму або мінімумів цільової функції. Ми обмежимося розглядом лише опуклих, ліпшицевих функцій, і більше того, ми побачимо, як навіть на простих прикладах результати оптимізації можуть варіюватися.

Почнемо із того, що розглянемо один із базових підходів, основу у даній роботі - метод градієнтного спуску. Цей метод є добре відомим і вивченим, але є багато варіацій методів, які ґрунтуються на його основі. Один з них - метод лінійного каплінгу. Його сенс - поєднання декількох різних методів у лінійну комбінацію для пришвидшення обчислень. Власне звідси і походить назва.

Для початку ми розглянемо дискретні варіації методів. Це означає, що ми матимемо початкову точку, а кожна наступна ітерація методу буде базуватись на результаті попередньої ітерації. У цій розділі ми також розглянемо метод дзеркального спуску, який доповнить вище згадані методи.

Наступний етап роботи - неперервні варіації методів. Ми спробуємо узагальнити наші вже отримані результати для дискретних варіантів у щось більше загальне, та щось, що залежить від часу. Вочевидь ми будемо розглядати системи диференціальних рівнянь.

У кінці роботи ми побачимо як методи вишкої оптимізації можуть використовуватись - у даній роботі на прикладі логістичної регресії.

2 Мета роботи

За мету у цьому дослідженні було покладено вивчення поведінки методів випуклої оптимізації і їх збіжності та швидкості залежно від характеристик оптимізованих об'єктів. Особлива увага приділяється порівнянню методів градієнтного спуску та лінійного каплінгу, яке базується не лише на теоретичних знаннях, а й на практичних завданнях.

Також було встановлено ціль продемонструвати застосування методів випуклої оптимізації в контексті задачі логістичної регресії і розкрити їхню ефективність та потенціал.

3 Розглянуті методи

3.1 Градієнтний метод

Щоб досягти оптимального розв'язку, в ітеративному процесі оптимізації часто застосовується підхід, який полягає у послідовному зменшенні значення цільової функції на кожній ітерації. Для визначення ефективності такого підходу зазвичай використовується гладка функція $f(x)$, яка формалізує цю задачу.

Будемо називати функція f ліпшицевою при виконанні наступних умов:

$$\|\nabla f(x) - \nabla f(y)\| \leq L(f)\|x - y\|, \quad \forall x, y \in \mathcal{Q} \quad (1)$$

Також, зазначимо

$$\forall y, f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \quad (2)$$

Використаємо ліпшицевість функції аби покращити крок алгоритму градієн-
тного спуску:

$$x_{k+1} \rightarrow \arg \min_y \left(\frac{L}{2} \|y - x_k\|^2 + \langle \nabla f(x), y - x_k \rangle \right) \quad (3)$$

Із припущення гладкості можна зробити наступний висновок:

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \|\nabla f(x_k)\|^2 \quad (4)$$

Загальний вигляд кроку методу градієнтного спуску:

$$\tilde{x} = \text{Grad}(x) := \arg \min_{y \in \mathcal{Q}} \left\{ \frac{L}{2} \|y - x\|^2 + \langle \nabla f(x), y - x \rangle \right\} \quad (5)$$

Із загальних віжкомостей з теорія гладкого опуклого програмування знаємо
такі оцінки:

$$f(\text{Grad}(x)) \leq f(x) - \text{Prog}(x) \quad (6)$$

Або, коли $\mathcal{Q} = \mathcal{R}^n$ $f(\text{Grad}(x)) \leq f(x) - \frac{1}{2L} \|\nabla f(x)\|^2$

Де $\text{Prog}(x) := - \min_{y \in \mathcal{Q}} \left\{ \frac{L}{2} \|y - x\|^2 + \langle \nabla f(x), y - x \rangle \right\} \geq 0$

Доведення: Нехай маємо, що $\tilde{x} = \text{Grad}(x)$. Тоді

$$\text{Prog}(x) := - \min_{y \in \mathcal{Q}} \left\{ \frac{L}{2} \|y - x\|^2 + \langle \nabla f(x), y - x \rangle \right\} = \quad (7)$$

$$- \left(\frac{L}{2} \|\tilde{x} - x\|^2 + \langle \nabla f(x), \tilde{x} - x \rangle \right) f(x) - \left(\frac{L}{2} \|\tilde{x} - x\|^2 + \langle \nabla f(x), \tilde{x} - x \rangle + f(x) \right) \quad (8)$$

$$\leq f(x) - f(\tilde{x}) \quad (9)$$

(7) - наслідок випуклості f . Також $\forall x, y \in \mathcal{Q}$ маємо наступне

$$\begin{aligned}
 f(x) - f(y) &= \int_0^1 \langle \nabla f(x + \tau(y - x)), y - x \rangle d\tau \\
 &= \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle d\tau \leq \\
 &\leq \langle \nabla f(x), y - x \rangle + \int_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|_* \cdot \|y - x\| d\tau \leq \\
 &\leq \langle \nabla f(x), y - x \rangle + \int_0^1 \tau L \|y - x\| \cdot \|y - x\| d\tau = \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2
 \end{aligned}$$

Наведемо доведення збіжності зазначеного методу:

$$\begin{aligned}
 \|x_{k+1} - x_k\|^2 &= \|x_k - \frac{1}{L} \|\nabla f(x_k)\| - x^*\|^2 = \\
 \|x_k - x^*\|^2 - \frac{2}{L} (\nabla f(x_k), x_k - x^*) + \frac{1}{L^2} \|\nabla f(x_k)\|^2
 \end{aligned}$$

Використовуючи, що $\nabla f(x^*) = 0$ матимемо

$$(\nabla f(x_k), x_k - x^*) = (\nabla f(x_k) - \nabla f(x^*), x_k - x^*) \geq \frac{1}{L} \|\nabla f(x_k)\|^2 \quad (10)$$

Завершуючи доведення маємо, що:

$$\begin{aligned}
 \|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - \frac{1}{L^2} \|\nabla f(x_k)\|^2 \Rightarrow \\
 \|x_k - x^*\|^2 &\searrow \\
 \frac{1}{L^2} \|\nabla f(x_k)\|^2 &\leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \Rightarrow \\
 \frac{1}{L^2} \sum_{k=0}^n \|\nabla f(x_k)\|^2 &\leq \|x_0 - x^*\|^2 \Rightarrow \\
 \sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2 &\leq +\infty \Rightarrow \nabla f(x_k) \rightarrow 0
 \end{aligned}$$

Доведення закінчено.

Наступний крок - отримати оцінку швидкості збіжності методу градієнтного

спуску:

Теорема: Якщо f та x_k задовольняють попереднім властивостям, матиме місце таке твердження:

$$f(x_T) - f(x^*) \leq O\left(\frac{LR^2}{T}\right), \quad R := \max_{x: f(x) \leq f(x_0)} \|x - x^*\|$$

Доведення: Із властивостей маємо

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

З опуклості на нерівності Коші-Шварца отримуємо:

$$\begin{aligned} f(x_k) - f(x^*) &\leq \langle \nabla f(x), x_k - x^* \rangle \leq \\ &\leq \|\nabla f(x_k)\| \|x_k - x^*\| \leq R \|\nabla f(x_k)\| \end{aligned}$$

Покладаючи $D_k = f(x_k) - f(x^*)$, матимемо дві нерівності:

$$D_k^2 \leq LR^2(D_k - D_{k+1}) \Rightarrow \frac{D_k}{D_{k+1}} \leq 2LR^2\left(\frac{1}{D_{k+1}} - \frac{1}{D_k}\right)$$

З того, що $D_k \geq D_{k+1}$, бо з кожним кроком ми наближуємось, ми отримаємо що:

$$\frac{1}{D_{k+1}} - \frac{1}{D_k} \geq \frac{1}{2LR^2}$$

З того, що на ітерації T маємо $\frac{1}{D_T} \geq \frac{T}{2LR^2}$, з чого матимемо нерівність $f(x_T) - f(x^*) \leq \frac{2LR^2}{T}$

3.2 Градієнтний метод для μ -сильно випуклих функцій

Вище було розглянуто метод градієнтного спуску для випуклої L -ліпшицевої функції. Якщо ми додамо умову μ -сильної випуклості. Таке обмеження дозволить прискорити роботу методу.

Означення: функція f називається μ -сильно випуклою, якщо $\exists \mu > 0$ таке що:

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) - \frac{1}{2}\mu\alpha(1 - \alpha)\|x - y\|^2 \quad (11)$$

Або $f - \frac{\mu}{2}\|\cdot\|^2$ - опукла функція.

Лема: Нехай f - неперервно диференційовна функція, тоді

$$f - \mu\text{-сильно випукла} \Leftrightarrow \\ \forall x, y \in \mathcal{Q} : f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2$$

Доведення:

$$(1 - \alpha)f(y) - (1 - \alpha)f(x) \geq f(\alpha x + (1 - \alpha)y) - f(x) + \frac{\mu}{2}\alpha(1 - \alpha)\|x - y\|^2 \Rightarrow \\ \Rightarrow f(y) - f(x) \geq \frac{f(\alpha x + (1 - \alpha)y) - f(x)}{1 - \alpha} + \frac{\mu}{2}\alpha\|x - y\|^2 \quad \alpha \rightarrow 1$$

Наведемо ще одну потрібну лему

Лема: f - неперервно диференційовна, μ -сильно опукла \Leftrightarrow

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu\|x - y\|^2 \quad (12)$$

Зазначимо наступне: сильно опукла функція на замкненій опуклій множині досягає мінімуму, причому точка мінімуму - єдина.

Нехай, f - μ -сильно випукла, а також $\exists x^*$ - точка мінімум. Тоді

$$f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2$$

Тепер, враховуючи μ -сильно випуклість функції f дамо наступну оцінку збіжності градієнтного методу

$$\begin{aligned}
f(x_{k+1}) - f(x_k) &\leq -\frac{1}{2L} \|\nabla f(x_k)\|^2 \\
f(x_{k+1}) - f(x^*) &\leq \frac{1}{2\mu} \|\nabla f(x_k)\|^2 \\
\Rightarrow 2\mu(f(x_k) - f(x^*)) &\leq \|\nabla f(x_k)\|^2 \leq 2L(f(x_k) - f(x_{k+1})) \\
\Delta_k &:= f(x_k) - f(x^*) \Rightarrow \\
\mu * \Delta_k &\leq L(\Delta_k - \Delta_{k+1}) \Rightarrow \Delta_{k+1} \leq (1 - \frac{\mu}{L})\Delta_k \Rightarrow \\
\Delta_k &\leq (1 - \frac{\mu}{L})\Delta_0 \leq e^{-\frac{\mu k}{L}} \Delta_0 \Rightarrow \\
\|x_k - x^*\|^2 &\leq \sqrt{\frac{2}{\mu}} e^{-\frac{\mu k}{2L}} \Delta_0^{\frac{1}{2}}
\end{aligned}$$

У роботі до сих пір розглядалися методи для параметра $\lambda = \frac{1}{L}$. Тепер спробуємо зробити щось подібне для $\lambda = \frac{1}{\mu+L}$

$$\begin{aligned}
x_{k+1} &= x_k - \frac{2}{\mu+L} \nabla f(x_k) \\
\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \frac{2}{\mu+L} \nabla f(x_k)\|^2 = \\
&\|x_k - x^*\|^2 - \frac{4}{\mu+L} \langle \nabla f(x_k), x_k - x^* \rangle + \frac{4}{(\mu+L)^2} \|\nabla f(x_k)\|^2 \leq \\
\leq^1 \|x_k - x^*\|^2 - \frac{4}{\mu+L} (\frac{\mu L}{\mu+L} \|x_k - x^*\|^2 + \frac{1}{\mu+L} \|\nabla f(x_k)\|^2) + \frac{4}{(\mu+L)^2} \|\nabla f(x_k)\|^2 & \\
&= (1 - \frac{4\mu L}{(\mu+L)^2}) \|x_k - x^*\|^2 \Rightarrow \\
\|x_{k+1} - x_k\|^2 &\leq (\frac{L-\mu}{L+\mu})^2 \|x_k - x^*\|^2 = (\frac{c-1}{c+1})^2 \|x_k - x^*\|^2 = \\
&= (1 - \frac{2}{c+1})^2 \|x_k - x^*\|^2 \Rightarrow \|x_k - x^*\|^2 \leq (1 - \frac{2}{c+1})^{2k} \|x_0 - x^*\|^2 \\
f(x_k) - f(x^*) &\leq \frac{L}{2} \|x_k - x^*\|^2 \leq \frac{L}{2} (1 - \frac{2}{c+1})^{2k} \|x_0 - x^*\|^2 \Rightarrow \\
f(x_k) - f(x^*) &\leq \frac{L}{2} e^{\frac{4ck}{c+1}} \|x_0 - x^*\|^2
\end{aligned}$$

4 Лінійний каплінг

4.1 Дзеркальний спуск

У цьому розділі ми розглянемо метод лінійного каплінгу. Проте перед тим, як увійти в деталі методу лінійного каплінгу, буде корисно розглянути метод дзеркального спуску (Mirror Descent). Для пояснення методу дзеркального спуску потрібно ввести кілька додаткових понять.

$$V_x(y) := w(y) - \langle \nabla w(x), y - x \rangle - w(x), \forall x \in Q \setminus \partial Q, \forall y \in Q$$

У означенні функція $w(\cdot)$ - 1-сильно опукла.

У роботі будуть розглянуті тільки такі функції у евклідовому просторі, де вона матиме наступний вигляд:

$$w(y) = \frac{1}{2} \|y\|^2, V_x(y) = \frac{1}{2} \|x - y\|^2$$

Тепер сформулюємо означення методу дзеркального спуску. Кроком методу дзеркального спуску довжини α назовемо:

$$\tilde{x} = \text{Mirr}_x(\alpha \partial f(x)), \text{ де } \text{Mir}_x(y) := \arg \min_{y \in Q} \{V_x(y) + \langle \alpha \partial f(x), y - x \rangle\}$$

Теорема (про дзеркальний спуск):

Якщо $x_{k+1} = \text{Mirr}_{x_k}(\alpha \partial f(x_k))$, то є правдою наступна нерівність:

$$\forall u \in Q, \alpha(f(x_k) - f(u)) \leq \alpha \langle \partial f(x_k), x_k - u \rangle \leq \frac{\alpha^2}{2} \|x_k\|^2 + V_{x_k}(u) - V_{x_{k+1}}(u)$$

Для доведення теореми спершу, необхідно показати істинність наступної рівності:

$$\langle \nabla V_x(y), y - x \rangle = V_x(u) - V_y(u) - V_x(y) \quad (13)$$

Дійсно:

$$\begin{aligned}
& \forall x, y \in \mathcal{Q} \quad \langle \nabla V_x(y), y - x \rangle = \langle \nabla w(x) - \nabla w(y), y - u \rangle \\
& = (w(u) - w(x) - \langle \nabla w(x), u - x \rangle) - (w(u) - w(y) - \langle \nabla w(y), u - y \rangle) \\
& \quad - (w(y) - w(x) - \langle \nabla w(x), y - x \rangle) \\
& = V_x(u) - V_y(u) - V_x(y)
\end{aligned}$$

Доведення: маємо наступні нерівності:

$$\begin{aligned}
& \alpha \langle \partial f(x_k), x_k - u \rangle = \langle \alpha \partial f(x_k), x_k - x_{k+1} \rangle + \langle \alpha \partial f(x_k), x_{k+1} - u \rangle \\
& \leq^1 \langle \alpha \partial f(x_k), x_k - x_{k+1} \rangle + \langle -\nabla V_{x_k}(x_{k+1}), x_{k+1} - u \rangle \\
& =^2 \langle \alpha \partial f(x_k), x_k - x_{k+1} \rangle + V_{x_k}(u) - V_{x_{k+1}}(u) - V_{x_k}(x_{k+1}) \\
& \leq^3 (\langle \alpha \partial f(x_k), x_k - x_{k+1} \rangle - \frac{1}{2} \|x_k - x_{k+1}\|^2) + (V_{x_k}(u) - V_{x_{k+1}}(u)) \\
& \leq^4 \frac{\alpha^2}{2} \|\partial f(x_k)\|^2 + (V_{x_k}(u) - V_{x_{k+1}}(u))
\end{aligned}$$

Покажемо правдивість нерівностей:

- 1) - З означенням $x_{k+1} = \arg \min_{x \in \mathcal{Q}} \{V_{x_k}(x) + \langle \alpha \partial f(x_k), x \rangle\}$, з чого випливає, що $\langle \nabla V_{x_k}(x_{k+1}) + \alpha \partial f(x_k), u - x_{k+1} \rangle \geq 0, \forall u \in \mathcal{Q}$
- 2) - безпосередньо випливає із рівності (13)
- 3) - наслідок з того, що $V_x(y) \geq \frac{1}{2} \|x - y\|^2$
- 4) - з теореми Коші-Шварца

Тепер, використовуючи розглянуті вище методи градієнтного та дзеркального спусків, сформулюємо крок методу лінійного каплінгу.

4.2 Алгоритм лінійного каплінгу

Input:

f – дифференційовна випукла, L -гладка функція

x_0 – якесь початкове значення

T – кількість операцій

for $k := 0$ **to** T **do**

$$\alpha_{k+1} \leftarrow \frac{k+2}{2L}$$

$$\tau_k \leftarrow \frac{2}{k+2}$$

$$x_{k+1} \leftarrow \tau_k z_k + (1 - \tau_k) y_k$$

$$y_{k+1} \leftarrow \text{Grad}(x_{k+1})$$

$$z_{k+1} \leftarrow \text{Mirr}(\alpha_{k+1} x_{k+1})$$

end

return y_T

Algorithm 1: Лінійний каплінг

Тепер наведемо основні твердження для збіжності наведеного алгоритму.

Теорема 1: Якщо $f(x)$ - L -гладка, випукла функція, $w(x)$ - 1-сильно випукла функція, тоді для алгоритму лінійного каплінгу має місце наступні нерівності:

$$f(y_T) - f(x^*) \leq \frac{4\Theta L}{T^2}, \text{ де } T - \text{кількість операцій.}$$

$$f(y_T) - f(x^*) \leq O\left(\frac{1}{T^2}\right)$$

Щоб довести зазначену теорему, наведемо декілька допоміжних тверджень:

Лема 1: Якщо $\tau_{k+1} = \frac{1}{\alpha_{k+1}L}$ тоді воно $\forall u \in \mathcal{Q}$ має місце наступна нерівність:

$$\begin{aligned} \alpha_{k+1} \langle \nabla f(x_{k+1}), z_k - u \rangle &\leq \alpha_{k+1}^2 L \text{Prog}(x_{k+1}) + V_{z_k}(u) - V_{z_{k+1}}(u) \leq \\ &\alpha_{k+1}^2 L (f(x_{k+1}) - f(y_{k+1})) + V_{z_k}(u) - V_{z_{k+1}}(u) \end{aligned}$$

Лема 2: $\forall u \in \mathcal{Q}$

$$(\alpha_{k+1}^2 L) f(y_{k+1}) - (\alpha_{k+1}^2 L - \alpha_{k+1}) f(y_k) + (V_{z_{k+1}}(u) - V_{z_k}(u)) \leq \alpha_{k+1} f(u)$$

Повернемось до доведення теореми 1. Використовуючи Лему 2, задаємо наступну послідовність α_k , таку що: $\alpha_k^2 L \approx \alpha_{k+1}^2 L - a_{k+1}$, а також $\tau = \frac{1}{\alpha_{k+1} L} \in (0; 1]$. В нащому методі каплінгу ми покладали, що $\alpha_k = \frac{k+1}{2L}$, тому $\alpha_k^2 L = \alpha_{k+1}^2 L - \alpha_{k+1} + \frac{1}{4L}$.

З Лемми 2 $k = 0, 1, \dots, T-1$ маємо наступне:

$$\alpha_T^2 L f(y_T) + \sum_{k=1}^{T-1} \frac{1}{4L} f(y_k) + (V_{z_T}(u) - V_{z_0}(u)) \leq \sum_{k=1}^T \alpha_k f(u)$$

Де беручи $u = x^*$ матимемо:

$$\sum_{k=1}^T \alpha_k = \frac{T(T+3)}{4L}, f(y_k) \geq f(x^*), V_{z_T}(u) \geq 0, V_{z_0}(u) \leq \Theta$$

Далі, використовуючи наведені значення маємо:

$$\frac{(T+1)^2}{4L^2} L f(y_T) \leq \left(\frac{T(T+3)}{4L} - \frac{T-1}{4L} \right) f(x^*) + \Theta$$

З чого слідує твердження теореми 1. Тепер перейдемо до доведення лем.

Лема 1: При $\tau_{k+1} = \frac{1}{\alpha_{k+1} L}$, то $\forall u \in \mathcal{Q}$ матимемо нерівність вигляду:

$$\begin{aligned} \alpha_{k+1} \langle \nabla f(x_{k+1}), z_k - u \rangle &\leq \alpha_{k+1}^2 L \text{Prog}(x_{k+1}) + V_{z_k}(u) - V_{z_{k+1}}(u) \leq \\ &\alpha_{k+1}^2 L (f(x_{k+1}) - f(y_{k+1})) + V_{z_k}(u) - V_{z_{k+1}}(u) \end{aligned}$$

Доведення: Друга нерівність отримана з методу градієнтного спуску, де ми отримали, що $f(x_{k+1}) - f(y_{k+1}) \geq \text{Prog}(x_{k+1})$. Для доведення першої нерівності перепишемо основну нерівність для методу дзеркального спуску.

$$\begin{aligned} \alpha_{k+1} \langle \nabla f(x_{k+1}), z_k - u \rangle &= \langle (\alpha_{k+1} \nabla f(x_{k+1}), z_k - z_{k+1}) \rangle + \langle \alpha_{k+1} \nabla f(x_{k+1}), z_{k+1} - u \rangle \\ &\leq^1 \langle \alpha_{k+1} \nabla f(x_{k+1}), z_k - z_{k+1} \rangle + \langle -\nabla V_{z_k}(z_{k+1}), z_{k+1} - u \rangle \\ &=^2 \langle \alpha_{k+1} \nabla f(x_{k+1}), z_k - z_{k+1} \rangle + V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1}) \\ &\leq^3 \left(\langle \alpha_{k+1} \nabla f(x_{k+1}), z_k - z_{k+1} \rangle - \frac{1}{2} \|z_k - z_{k+1}\|^2 \right) + (V_{z_k}(u) - V_{z_{k+1}}(u)) \end{aligned}$$

Де:

1) - випливає з того, що $z_{k+1} = \arg \min_{z \in (Q)} \{V_{z_k}(z) + \langle \alpha_{k+1} \nabla f(x_{k+1}), z \rangle\}$, виходячи з чого:

$$\langle \nabla V_{z_k}(z_{k+1}) + \alpha_{k+1} \nabla f(x_{k+1}), u - z_{k+1} \rangle \geq 0 \quad \forall u \in Q.$$

2) випливає із рівності Брегмана (була використана для доведення методу дзеркального спуску).

3) слідує з того, що $V_x(y) \geq \frac{1}{2} \|x - y\|^2$, що є наслідком сильної випуклості w . Далі буде використана нерівність Коші-Шварца: $\langle \alpha_{k+1} \nabla f(x_{k+1}), z_k - z_{k+1} \rangle - \frac{1}{2} \|z_k - z_{k+1}\|^2 \leq \frac{\alpha_{k+1}^2}{2} \|\nabla f(x_{k+1})\|^2$, то якраз отримаємо потрібну нерівність, зазначивши, що це правда для випадку коли $Q = \mathcal{R}^n$, бо $Prog(x_{k+1}) = \frac{1}{2L} \|\nabla f(x_{k+1})\|^2$ (з 7).

Лема 2: $\forall u \in Q$

$$(\alpha_{k+1}^2 L) f(y_{k+1}) - (\alpha_{k+1}^2 L - \alpha_{k+1}) f(y_k) + (V_{z_{k+1}}(u) - V_{z_k}(u)) \leq \alpha_{k+1} f(u)$$

Доведення: Маємо наступний ланцюжок нерівностей:

$$\begin{aligned} & \alpha_{k+1} (f(x_{k+1}) - f(u)) \\ & \leq \alpha_{k+1} \langle \nabla f(x_{k+1}), x_{k+1} - u \rangle \\ & = \alpha_{k+1} \langle \nabla f(x_{k+1}), x_{k+1} - z_k \rangle + \alpha_{k+1} \langle \nabla f(x_{k+1}), z_k - u \rangle \\ & =_1 \frac{(1 - \tau_k) \alpha_{k+1}}{\tau_k} \langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle + \alpha_{k+1} \langle \nabla f(x_{k+1}), z_k - u \rangle \\ & \leq_2 \frac{(1 - \tau_k) \alpha_{k+1}}{\tau_k} (f(y_k) - f(x_{k+1})) + \alpha_{k+1} \langle \nabla f(x_{k+1}), z_k - u \rangle \\ & \leq_3 \frac{(1 - \tau_k) \alpha_{k+1}}{\tau_k} (f(y_k) - f(x_{k+1})) + \alpha_{k+1}^2 L (f(x_{k+1}) - f(y_{k+1})) + V_{z_k}(u) - V_{z_{k+1}}(u) \\ & =^4 (\alpha_{k+1}^2 L - \alpha_{k+1}) f(y_k) - (\alpha_{k+1}^2 L) f(y_{k+1}) + \alpha_{k+1} f(x_{k+1}) + V_{z_k}(u) - (V_{z_{k+1}}(u)) \end{aligned}$$

Обґрунтуємо наведені переходи :

1) - випливає з вибору $x_{k+1} : \tau_k(x_{k+1} - z_k) = (1 - \tau_k)(y_k - x_{k+1})$

2) - в наслідок опуклості f і $1 - \tau_k \geq 0$

3) - застосовуючи лему 1,

4) - та з вибору $\tau_k = \frac{1}{\alpha_{k+1}L}$

Далі будуть розглянуті практичні приклади.

5 Приклади

Як вже було зазначено раніше, основною метою цієї роботи є не тільки надання визначення методу лінійного каплінга та теоретичне доведення його збіжності, але й дослідження його практичних застосувань та порівняння зі звичайним градієнтним спуском. На основі отриманих даних буде зроблений висновок щодо ефективності цього методу.

У цьому розділі будуть розглянуті прості приклади застосування методів опуклої оптимізації. Робота буде проводитися у евклідовому просторі, а функції, які будуть мінімізовані, можуть бути двовимірними або навіть тривимірними.

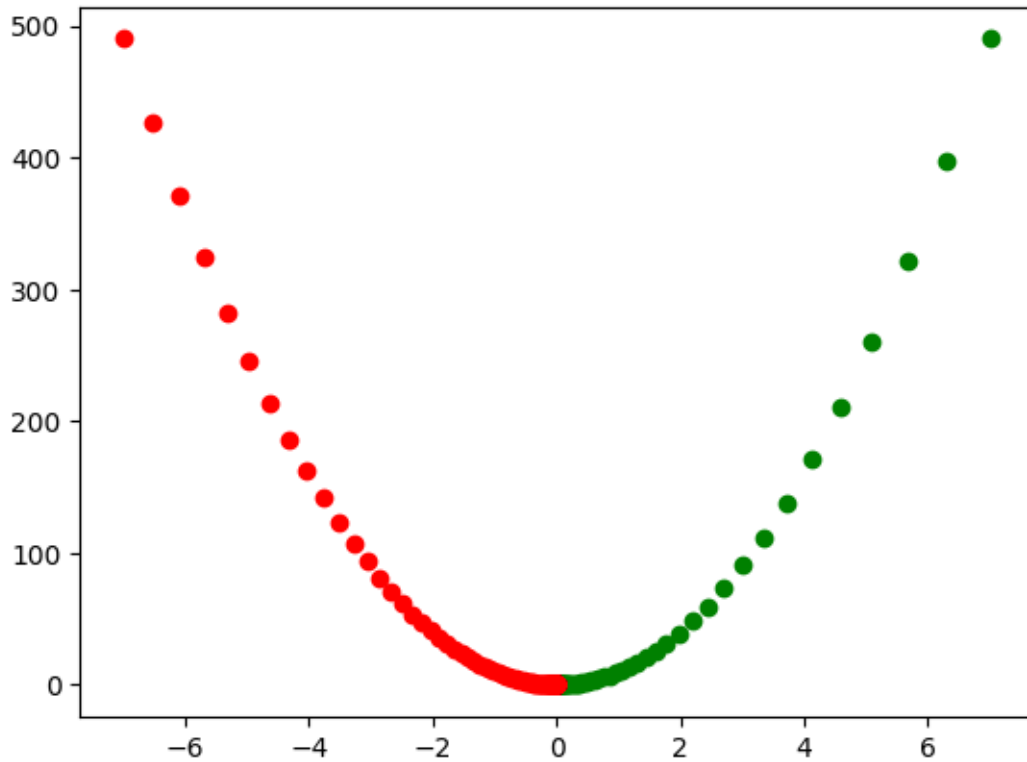
5.1 Розглянка Функція: $f = \frac{x^2 * n}{2}$

Перший приклад - для функції $f = \frac{x^2 * n}{2}$. Ясно видно, що мінімум функції буде досягнутий у точці 0. На цьому прикладі буде спроба аналізу різниці між методом градієнтного спуску з різними коефіцієнтами λ .

Отже наші методи виглядатимуть наступним чином:

$$\begin{aligned} 1) x_{k+1} &= x_k - \frac{1}{L} \nabla f(x_k) \\ 2) x_{k+1} &= x_k - \frac{1}{L + \mu} \nabla f(x_k) \end{aligned}$$

Проведемо дослідження швидкості збіжності наведених методів із наступними параметрами $\mu = 10$, кількість ітерацій $n = 1000$, $x_0 = 7$:



Номер ітерації	$x_k - \frac{1}{L} \nabla f(x_k)$	$x_k - \frac{1}{L+\mu} \nabla f(x_k)$
30	0.2967381079265133	-0.8834893907033995
50	0.0125790720994010077	-0.11150764335506627
99	0.000206588658014569	-0.007563780647430592
500	9.254495736365615e-23	-7.302748619381937e-15
1000	1.223509876205848e-45	-7.6185910568407e-30

Із наведених даних одразу видно різницю роюоти цих двох методів, вона не така вражаюча. Проте, як і раніше, дуже цікавим завданням є порівняння методів градієнтного спуску та лінійного каплінгу для цього ж прикладу. Після виконання програми було отримано наступні результати::

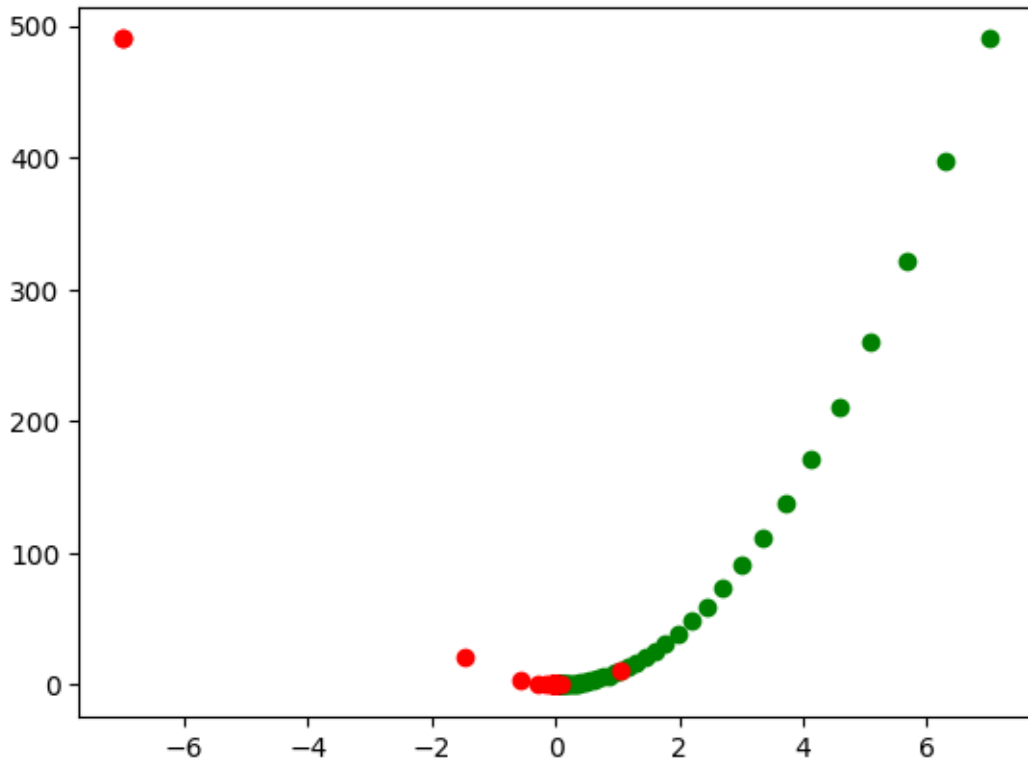


Рис. 1: Червоний - ліній каплінг, зеленій - градієнт

Номер ітерації	<i>Grad</i>	<i>LinearCoupling</i>
30	0.2967381079265133	-0.0034322164325116676
50	0.012579072099401007	-4.881342824203231e-05
99	0.000206588658014569	-3.055620631430923e-07
500	9.254495736365615e-23	-1.7473789360408612e-27
1500	1.223509876205848e-45	-4.515305578098292e-52

Як ми бачимо у раі порівняння навіть на цьому прикладі лінійни каплінг працює набагато швидше ніж градієнтний спуск. Також ми пересвідчились, що для цього варіанту f параметр n майже не має ніякого впливу на швидкість збіжності кожного з методів.

Щоб отримати саме видимо залежність швидкості методу від параметру функції, слід звернути увагу на приклад, розглянутий у наступній секції.

Із наведених результатів можна зрозуміти висновок, що для цього конкретного прикладу лінійний каплінг працює значно швидше, ніж градієнтний спуск. Також було продемонстровано, що для даного варіанту функції f параметр n майже не має впливу на швидкість збіжності обох методів.

У наступному розділі ми більш детально розглянемо різницю роботи наших методів, уже для тривірмного прикладу.

5.2 Розглянка Функція: $f(x, y) = \frac{1}{2} \frac{x^2 + ny^2}{n+1}$

На цьому прикладі видно, що зі збільшенням параметру функції оптимізація стає все складнішою. З метою порівняння методів, було протестовано обидва методи з різних початкових точок, використовуючи симетричність функції. Після виконання програми отримуємо наступні результати:

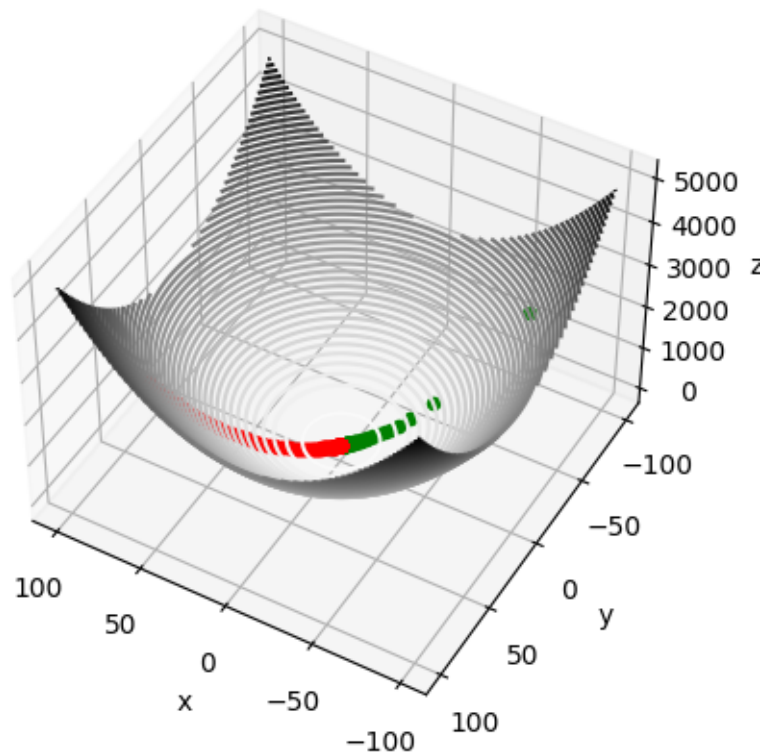


Рис. 2: Зелений - ліній каплінг, червоний - градієнт

Початковою точкою методу градієнтного спуску x_0 була $[70, 70]$, а методу лінійного каплінгу x_0 $[-70, -70]$. Параметр функції n покладено рівним 1.

Номер ітерації	<i>Grad</i>	<i>LinearCoupling</i>
30	[15.5478, 15.5478]	[-2.90012, -2.90012]
50	[4.79895, 4.79895]	[-7.9965-17, -7.9965e-17]
99	[0.590763, 0.590763]	[-9.1236-26, -9.1236-26]
500	[9.4567-10, 9.4567e-10]	[-9.194-133, 9.194-133]
1000	[5.9056e-32, 5.9056e-32]	[-2.639-285, -2.639-285]

На наведеному прикладі видно, що навіть у даному випадку метод лінійного каплінгу продемонстрував значно швидшу збіжність. Вже на 99-ій ітерації його результат був точнішим, ніж у градієнтного спуску на 500-ій ітерації. Для подальшого порівняння, змінимо параметр і встановимо значення $n = 20$. Виконавши програму з новим параметром, ми отримаємо наступний результат:

Результати на контрольних ітераціях:

Номер ітерації	<i>Grad</i>	<i>LinearCoupling</i>
30	[60.660756, 2.5260424]	[-41.712811, -1.167222]
50	[52.567524, -0.824887]	[-25.8659480, -1.062927]
99	[43.638538, -0.996466]	[-14.0311788, -1.025948]
500	[6.435669, -0.999994]	[-0.0350511, -1.000091]
1000	[0.59168, -0.999994]	[-2.55033e-05, -1.00009]

Зауважимо, що зі збільшенням параметра n до 20 ми отримали значну різницю в роботі обох методів. Як і раніше, лінійний каплінг виявився швидшим і показав більш точний результат. Це підтверджує його переваги у порівнянні з градієнтним спуском для даного прикладу.

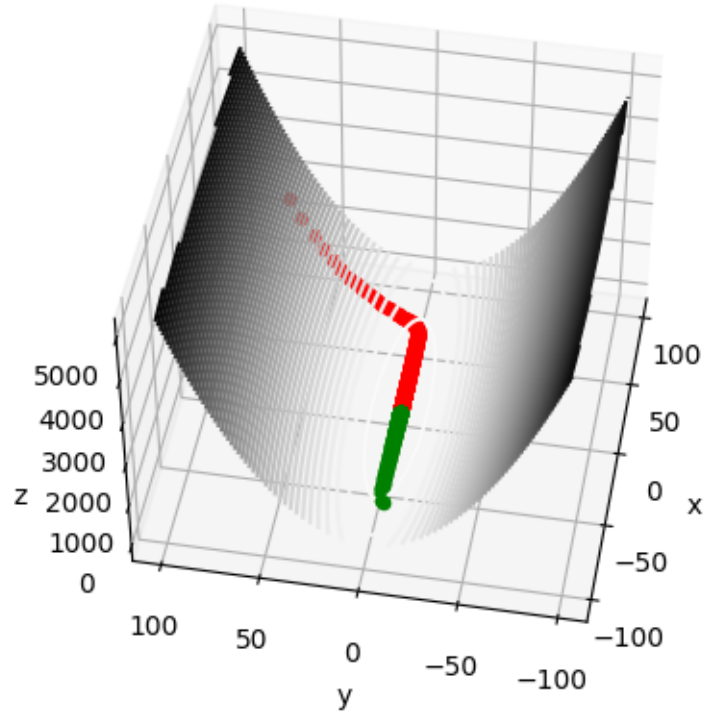


Рис. 3: Зелений - ліній каплінг, червоний - градієнт

Основні висновки, які можна зробити з проведених тестів, такі:

1. Метод лінійного каплінгу збіжний і є валідним для розв'язання проблеми оптимізації у принаймні у евклідовому просторі.
2. Швидкість роботи методу лінійного каплінгу перевищує швидкість градієнтного спуску незалежно від розглянутого прикладу.
3. Обидва методи, що були протестовані, суттєво залежать від параметрів функції. Навіть невелика зміна параметрів може суттєво вплинути на швидкість збіжності методів.

6 Неперервний випадок

У цьому розділі ми розглянемо, власне, основне питання цієї роботи, тобто введемо означення методам у неперервному випадку, покажемо їх теоретичну збіжність і порівняємо їх роботу на практиці. Варто зазначити, що у багатьох випадках у подальшому ми будемо користуватися фактами, які невдовзі будуть доведені. Перш за все давайте введемо декілька нових позначень.

Нехай маємо рівняння вигляду

$$\begin{cases} \dot{x} = \phi(x) \\ x(0) = x_0 \end{cases}$$

Розглянемо функцію $V : R^n \rightarrow R, V \in C^1(R^n)$ таку що $V(x) \geq V^*$

Покладемо $t \rightarrow x(t), f = V(x(t))$, тоді:

$$\dot{V}(x, t) = \langle \nabla V(x(t)), \dot{x}(t) \rangle = \langle \nabla V(x(t)), \phi(x(t)) \rangle$$

$\dot{V}(x) = \langle \nabla V(x), \phi(x) \rangle$ - похідна в силу нашої системи. Також введемо функцію $W(x) \in C(R^n)$ таку що : $-\dot{V}(x) \geq W(x) \geq 0, \forall x \in R^n$. Слід зазначити, що на цю теорему буде посылатись велика кількість інших фактів, які згодом будуть зустрічатися у цій роботі, тому на неї слід звернути більше уваги.

Теорема: мають місце наступні твердження:

1) $\lim_{t \rightarrow \infty} W(x(t)) = 0, \min_{0 \leq t \leq T} W(x(t)) \leq \frac{V(x(0)) - V^*}{T}$

2) Нехай W - опукла, а $\tilde{x}(T) = \frac{1}{T} \int_0^T x(t) dt$, тоді:

$$W(\tilde{x}(T)) \leq \frac{V(x(0)) - V^*}{T}$$

3) Якщо:

а) $V(x^*) = V^*, V(x) > V^*, \forall x \neq X^*$

б) $\inf_{\|x - x^*\| \geq R} (V(x) - V^*) \geq \epsilon > 0$

в) $W(x^*) = 0, W(x) > 0, x \neq x^*$, то

$$x(t) \rightarrow x^*, t \rightarrow \infty$$

4) $\exists \gamma > 0$, і якщо $W(x) \geq \gamma(V(x) - V^*) \Rightarrow$

$$V(x(t) - V^*) \leq (V(x(0) - V^*) * \exp -\gamma t$$

Доведення:

$$\begin{aligned} 1) W(x(t)) \leq V(\dot{x}(t)) &\Rightarrow \int_0^T W(x(t)) dt \leq - \int_0^T \langle \nabla V(x(t)), \phi(x(t)) \rangle \\ &= V(x(0)) - V(T) \leq V(x(0)) - V^* \end{aligned}$$

$$\begin{aligned} \forall T > 0 \min_{t \in [0; T]} &\leq \int_0^T W(x(t)) dt \leq V(x(0)) - V^* \Rightarrow \\ \Rightarrow \int_0^\infty W(x(t)) dt < +\infty &\Rightarrow \lim_{t \rightarrow \infty} W(x(t)) = 0 \end{aligned}$$

2) використовуючи інтегральну нерівність Єнсена маємо:

$$W\left(\frac{1}{T} \int_0^T x(t) dt\right) \leq \frac{1}{T} \int_0^T W(x(t)) dt \leq \frac{1}{T} (V(x(0)) - V^*)$$

3) - класична форма теореми Ляпонова про асимптотичну стабільність*

$$\begin{aligned} 4) \quad W(x) \geq \gamma(V(x) - V^*) &\Rightarrow -\dot{V} \geq \gamma(V(x) - V^*) \quad | * e^{\gamma t} \\ &\Rightarrow \gamma e^{\gamma t} * (V(x) - V^*) + e^{\gamma t} \dot{V} \leq 0 \Rightarrow \\ \Rightarrow \int_0^t \frac{d}{dt} (e^{\gamma t} (V(x) - V^*)) dt &= (e^{\gamma t} (V(x(t)) - V^*) - (V(x(0)) - V^*)) \leq 0 \\ \Rightarrow V(x(t)) - V^* &\leq e^{-\gamma t} (V(x(0)) - V^*) \end{aligned}$$

Доведення закінчено.

Тепер дамо означення методу градієнтного спуску для неперервного випадку, $f - \mu$ -сильно опукла

$$\begin{cases} \dot{x} = -\nabla f(x) \\ x(0) = x_0 \end{cases}$$

$$V(x) = \frac{1}{2}\|x - x^*\|^2, f(x^*) = f^*$$

$$\begin{aligned} \dot{V}(x(t)) &= \langle \dot{x}, x - x^* \rangle = -\langle \nabla f(x), x - x^* \rangle \leq \\ &\leq^1 -\mu\|x - x^*\|^2 = -2\mu V(x) \Rightarrow \dot{V}(x) + 2\mu V(x) \leq 0 \quad | * e^{2\mu t} \\ \frac{d}{dt}(e^{2\mu t} V) &\leq 0 \Rightarrow \int_0^t \frac{d}{dt} e^{2\mu t} V dt = e^{2\mu t} V - V(x_0) \leq 0 \Rightarrow \\ &\Rightarrow V(x(t)) \leq^2 e^{-2\mu t} V(x_0) \Rightarrow \\ \|x(t) - x^*\|^2 &\leq e^{-2\mu t} \|x_0 - x^*\|^2 \Rightarrow \\ \Rightarrow^3 f(x(t)) - f^* &\leq \frac{L}{2} e^{-2\mu t} \|x_0 - x^*\|^2 \end{aligned}$$

Де:

- 1) - з μ -сильно опуклості f
- 2) - нерівність Гронуолла
- 3) - з леми спуску ($f(x) - f^* \leq \frac{L}{2}\|x - x^*\|^2$)

6.1 Додаткові теоретичні результати

Нам необхідно також показати істинність ще декількох тверджень аби перейти безпосередньо до методів. Із теореми з минулої секції ми можемо отримати наступні висновки:

- 1) Якщо $f(x)$ обмежена знизу: $f(x) \geq f^*$, то

$$\min_{0 \leq t \leq T} \|f'(x(t))\|^2 \leq \frac{f(x(0)) - f^*}{T}$$
- 2) Якщо до того ж має місце нерівність

$$\|f'(x)\|^2 \geq 2\mu(f(x) - f^*), \mu > 0, \text{ то}$$

$$f(x(t)) - f^* \leq ((f(x(0)) - f^*)e^{-2\mu t})$$

3) Якщо $f(x)$ випукла та існує x^* такий що $f(x^*) = f^*$, то

$$f(\bar{x}(T)) - f^* \leq \frac{\|x(0) - x^*\|^2}{T}$$

4) Якщо $f(x)$ сильно випукла тоді $x(t) \rightarrow x^*$ експотенційно, при $t \leftarrow \inf$

Доведення:

Для доведення перших двох тверджень покладемо функцію $V : V(x) = f(x)$.

В такому випадку маємо, що $W(x) = \|f'(x)\|^2$. Тоді використовуємо твердження 1) та 4) із передньої теореми.

Для доведення 3) та 4) тверджень змінимо нашу функцію V , покладаючи $V(x) = \frac{1}{2}\|x - x^*\|^2$. Тоді відповідно матимемо $W(x) = f(x) = f^*$. Тепер із 2) твердження передньої теореми випливає 3). В той же час для сильно випуклих функцій існує точка мінімуму x^* , звідки і випливає твердження 4).

Теорема доведена.

Зазначимо, що для твердження 4) не випливає єдиність такої точки мінімуму.

Наприклад, якщо $g(x) = R^n \rightarrow R^m, f(x) = \|g(x)\|^2$, тоді матимемо, що $f(x)$ не є опуклою але 4) твердження виконується.

6.2 Метою важкої кульки. Загальний вигляд

Отже тепер дамо визначимо методу важкої кульки у загальному випадку. метод важкої кульки має наступний вигляд:

$$\ddot{x} + a\dot{x} + bf'(x) = 0$$

Такий вигляд був природнім результатом із аналогії із фізики: рух тіла в потенціальному полі $f(x)$ піддається тертю. Тут $x(t) \in R^n$ – траєкторія тіла, \dot{x} , \ddot{x} – його швидкість і прискорення, $a, b > 0$ є скалярними параметрами. Природно

припустити, що, при $t \rightarrow \inf$, тіло досягає локального мінімуму $f(x)$.

Давайте тепер перепишемо наше початкове рівняння у наступному вигляді:

$$\begin{aligned}\dot{x} &= y, \\ \dot{y} &= -ay - bf'(x)\end{aligned}$$

Тепер маємо багато опцій у виборі функції $V(x)$. Природно, із фізичних міркувань покласти за $V(x) = f(x) + \frac{1}{2b}\|y\|^2$ - сума потенціалів та кінетичної енергії. Похідна від цієї функції матиме наступний вигляд:

$$\begin{aligned}\dot{V}(x, y) &= (f'(x), y) + \frac{1}{2b}(y, -ay - bf'(x)) = -\frac{a}{b}\|y\|^2 \\ W &= \frac{a}{b}\|y\|^2 > 0\end{aligned}$$

Але ми б хотіли побудувати строго спадну функцію Ляпунова, тому матимемо наступне:

$$V = (f(x) - f^*) + \frac{a}{a^2 + 2bL}(f'(x), y) + \frac{L}{a^2 + 2bL}\|y\|^2$$

Де f^* - нижня границя $f(x)$, L - верхня границя $\|f''(x)\|$. Тепер ми можемо застосувати попередню теорему і отримати наступні результати:

Теорема: Нехай $f(x)$ двічі диференційовна функція, де $\|f''(x)\| < L$, $f(x) \geq f^*$ та $a, b > 0$. Тоді матиме місце наступна нерівність:

$$\min_{0 \leq t \leq T} \|f'(x(t))\| \leq \frac{V(0) - f^*}{cT}, c = \frac{ab}{a^2 + 2bL}$$

Доведення: Перш за все пригадаємо, що для нашої функції справджується умова

$$f(x) - f \geq \frac{1}{2L}\|f'(x)\|^2$$

Тому матимемо наступне:

$$\begin{aligned} V &\geq \frac{1}{2L} \|f'(x)\|^2 + \frac{a}{a^2 + 2bL} (f'(x), y) + \frac{L}{a^2 + 2bL} \|y\|^2 \\ &\geq c_1 (\|f'(x)\|^2 + \|y\|^2), c_1 > 0 \end{aligned}$$

Далі, обчислюючи похідну маємо:

$$\dot{V} \leq \frac{La}{a^2 + 2bL} \|y\|^2 - \frac{ab}{a^2 + 2bL} \|f'(x)\|^2$$

З чого слідувало б наступне:

$$W \geq c_2 (\|f'(x)\|^2 + \|y\|^2), c_2 > 0$$

Тепер використовуючи твердження 1) з попередньої теореми отримуємо наше твердження.

Зазначимо також, що для випуклих $f(x)$ можливий і інший вигляд функції Ляпунова. Наприклад

$$V = (f(x) - f^*) + \frac{a^2}{2b} \|x - x^* + \frac{1}{a} y\|^2$$

Теорема: якщо $f(x)$ опукла та $f(x^*) = f^*$ тоді для методу важкої кульки матимемо, що:

$$\begin{aligned} \bar{x}(T) &= \frac{1}{T} \int_0^T x(t) dt \\ f(\bar{x}(T)) - f^* &\leq \frac{V(x(0), y(0))}{aT} \end{aligned}$$

Доведення: для нашої функції обчислюємо похідну:

$$\dot{V} = -a(f'(x), x - x^*)$$

Також покладемо $W(x) = a(f(x) - f^*)$. Тепер можемо використати твердження попередніх теорем, із чого слідуватиме твердження.

Варто зазначити, що при $a = \frac{c}{t}$ та із функцією Ляпунова вигляду $V = (f(x) - f^*) + \frac{a^2}{2b} \|x - x^* + \frac{1}{a}y\|^2$ метод важкої кульки збігається зі швидкістю $O(\frac{1}{t^2})$

6.3 Метод важкої кульки у гамільтоновій формі.

У цій секції ми розглянемо дещо загальніший вигляд методу важкої кульки у так званій гамільтоновій формі.

Перш за все, нехай ми маємо рівняння вигляду

$$\ddot{x} + a\dot{x} + b \sin x - c = 0, a > 0, 0 < b < c, x(t) \in R^1$$

Нас цікавить перевірка стабільності точки рівноваги, тобто $x = x^* = \arcsin \frac{c}{b}, y = \dot{x} = 0$. Традиційно для аналізу асимптотичної стабільності ми маємо таку функцію:

$$V = \frac{y^2}{2} + b(\cos x^* - \cos x) - c(x - x^*)$$

Для такої функції ми маємо, що $\dot{V} = -W = -ay^2 \leq 0$ і для доведення стабільності використовується результат теореми Барбашина—Красовського. Але замість цього давайте задамо функцію Ляпунова у наступному вигляді:

$$V(x, y) = \frac{y^2}{2} + \frac{ap}{2}(x - x^*)^2 + py(x - x^*) + b(\cos x^* - \cos x) - c(x - x^*)$$

Тут p -параметр, який розглянемо трохи пізніше, $V(x^*, 0) = 0, V(x, y) > 0$ у околі точки x^* . Варто зазначити, що показати асимптотичну стабільність можна буде лише у локальному околі точки x^* .

Теорема: Якщо $0 < p < a, |x(0) - x^*| < \pi$ та

$$V(x(0), y(0)) < h = V(\pi - x^*, -p(\pi - x^*))$$

Тоді $\ddot{x} + a\dot{x} + b \sin x - c = 0$, $a > 0$, $0 < b < c$, $x(t) \in R^1$ локально асимптотично стабільна система.

Доведення: Для початку обчислимо похідні наших функцій:

$$\dot{V} = -(a - p)y^2 - bp(x - x^*)(\sin x - \sin x^*)$$

$$W = (a - p)y^2 + bp(x - x^*)(\sin x - \sin x^*)$$

Також відмітимо, що з умов $0 < p < a$ та $|x + x^*| < \pi$ слідує, що $W \geq 0$ та $W > 0$ для $(x, y) \neq (x^*, 0)$. Далі покажемо, що $V \geq 0$ для

$$(x, y) \in Q = \{x : |x + x^*| < \pi\}$$

Позначимо $v(x) = \min_y V(x, y)$, тоді

$$y(x) = \operatorname{argmin}_y V(x, y) = -p(x - x^*),$$

$$v(x) = \frac{(a - p)p}{2}(x - x^*)^2 + b(\cos x^* - \cos x) - c(x - x^*)$$

Тоді відповідно маємо

$$v'(x) = (a - p)p(x - x^*) + b(\sin x - \sin x^*)$$

$$v(x) > 0 : |x + x^*| < \pi, (x, y) \neq (x^*, y)$$

З цього можемо зробити висновок, що $V(x, y) > 0$, якщо $|x + x^*| < \pi$, $(x, y) \neq (x^*, y)$. Більше того множина

$$D = \{(x, y) : V(x, y) < h, |x - x^*| < \pi\}$$

Обмежена та належить Q . Тепер можемо застосувати попередню теорему.

Такий підхід ми можемо розширити для загального випадку у наступному вигляді:

$$\ddot{x}_k + a_k \dot{x}_k + \sum_j b_{jk} \sin(x_k - x_j) - c_k = 0, x_k(t) \in R^1$$

Тоді ми можемо переписати це у наступному вигляді:

$$\dot{x} = Ax + B\varphi(Cx)$$

Де A, B, C - матриці, а $\varphi(z)$ - нелінійний вектор. Тоді функція Ляпонова матиме наступний вигляд:

$$V(x) = (Px, x) + K\psi(Cx)$$

Де потрібно обрати матриці P, K , а $\psi'(x) = \varphi(x)$ (z). Лінійні матричні нерівності для P, K були отримані в (Vu and Turitsyn, 2016). Цей розв'язок власне і визначає шукані функції V, V' із потрібним властивостями.

7 Приклад з логістичною регресією

7.1 Приклад для двовимірного розділення точок

В цьому розділі прикладів ми розглянемо випуклу оптимізацію методами градієнтного спуску та лінійного каплінгу для задачі логістичної регресії. Де-кілька слів про саму задачу:

Нехай маємо деяку випадкову величину Y , що може набувати двох значень (два кольори, 0 та 1, тощо). Для зручності вважаємо просто, що $Y \in \{0, 1\}$.

Нехай, також, ця випадкова величина залежить від деякої множини пояснювальних змінних $x \in \mathcal{R}^n$, де

$$x = (x_1, x_2, \dots, x_n)^T$$

Для того, щоб показати залежність Y від $x^{(i)}$ -их введемо додаткову змінну y^* :

$$y^* = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

Тоді задати нашу Y буде значно зручніше, і виглядатиме вона приблизно так:

$$Y = \begin{cases} 0 : y^* \leq 0 \\ 1 : y^* > 0 \end{cases}$$

При визначенні логістичної моделі стохастичний доданок ϵ вважається випадковою величиною з логістичним розподілом ймовірностей. Відповідно для певних конкретних значень змінних $x^* = x_1^*, x_2^*, \dots, x_n^*$ одержується відповідне значення y^* і ймовірність того, що $Y = 1$, така:

$$\begin{aligned} p(Y = 1) &= p(y^* > 0) = p(\theta^T x^* + \epsilon > 0) = p(\epsilon > -\theta^T x^*) = \\ &= p(\epsilon \leq \theta^T x^*) = g(\theta^T x^*) \end{aligned}$$

Передостання рівність впливає з симетричності логістичного розподілу, g позначає логістичну функцію - функцію розподілу:

$$g(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

Таким чином для конкретного значення x^i випадкова величина Y^i , має розподіл Бернуллі:

$$Y^i = \mathcal{B}(1, g(\theta^T x^i))$$

Тепер наше завдання: оцінити параметри $\theta_0, \theta_1, \dots, \theta_n$ так, щоб зв'язок між Y від $x^{(i)}$, був як можливо ближчий до істини.

Таку оцінку зазвичай здійснюють методом максимальної правдоподібності, згідно з яким вибирають параметри θ , що максимізують значення функції правдоподібності на вибірці:

$$\hat{\theta} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \prod_{i=1}^n P\{Y = Y^{(i)} | x = x^{(i)}\}$$

Максимізація функції правдоподібності еквівалентна максимізації її логарифма:

$$\ln L(\theta) = \sum_{i=1}^n \ln P\{Y = Y^{(i)} | x = x^{(i)}\} = \sum_{i=1}^n Y^{(i)} \ln(g(\theta^T x^{(i)})) + (1 - Y^{(i)}) \ln(1 - g(\theta^T x^{(i)}))$$

Наші методи градієнтного спуску та лінійного каплінга ми будемо застосовувати саме для пошуку екстремальних значення цієї функції.

Перевіряти збіжність наших методів будемо на наступному прикладі, побудованому на цій задачі.

1. На вхід користувач вводить за допомогою мишки певну кількість точок різного кольору (у програмі значення $Y^{(i)} = 1$ відповідає зеленому кольору, відповідно значення 0 - червоному)
2. Після деякої кількості точок (вона визначена заздалегідь), на основі отриманих даних ми використаємо наш метод для оптимізації функції правдоподібності
3. Після роботи методу ми отримаємо наш вектор θ (оскільки в нас приклад на площині, то $\theta \in R^3$)
4. На базі отриманого вектора будемо пряму, яка задана рівняння $\theta_0 + \theta_1 x + \theta_2 y = 0$. Саме ця пряма і буде зображена на графіку

Для роботи будь-якого з описаних методів нам перш за все потрібно знайти градієнт від цільової функції. У підрахунках може з'явитися проблема: якщо

$$y^* = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n,$$

то e^{y^*} може бути дуже великим, в результаті чого ми можемо отримати переповнення та програма не буде повністю валідною (проте все одно, напевне, залишається можливість підбрати початкові значення та початкові точки так, щоб переповнення не було).

Враховуючи це ми будемо брати градієнт наступним чином:

$$\begin{aligned}
 & \text{if}(y^*) > 0 \quad \text{then} \\
 & \theta_0 = \sum_{i=1}^m y^{(i)} * \left(\frac{e^{-y^*}}{1 + e^{-y^*}} \right) + (1 - y^{(i)}) * \left(\frac{-1}{1 + e^{-y^*}} \right) \\
 & \theta_k = \sum_{i=1}^m y^{(i)} * \left(\frac{x_1^{(k)} e^{-y^*}}{1 + e^{y^*}} \right) + (1 - y^{(i)}) * \left(-x_1^{(k)} - \frac{x_1^{(k)} * e^{-y^*}}{1 + e^{-y^*}} \right) \\
 & \text{else}(y^*) \leq 0 \\
 & \theta_0 = \sum_{i=1}^m y^{(i)} * \left(\frac{1}{1 + e^{y^*}} \right) + (1 - y^{(i)}) * \left(1 - \frac{1}{1 + e^{y^*}} \right) \\
 & \theta_k = \sum_{i=1}^m y^{(i)} * \left(\frac{x_k^{(i)}}{1 + e^{y^*}} \right) + (1 - y^{(i)}) * \left(-x_k^{(i)} * \frac{e^{y^*}}{1 + e^{y^*}} \right)
 \end{aligned}$$

Таке визначення складових градієнта, як можна помітити, гарантує, що у нас не буде переповнення і всі змінні вдається без зайвих клопотів обрахувати.

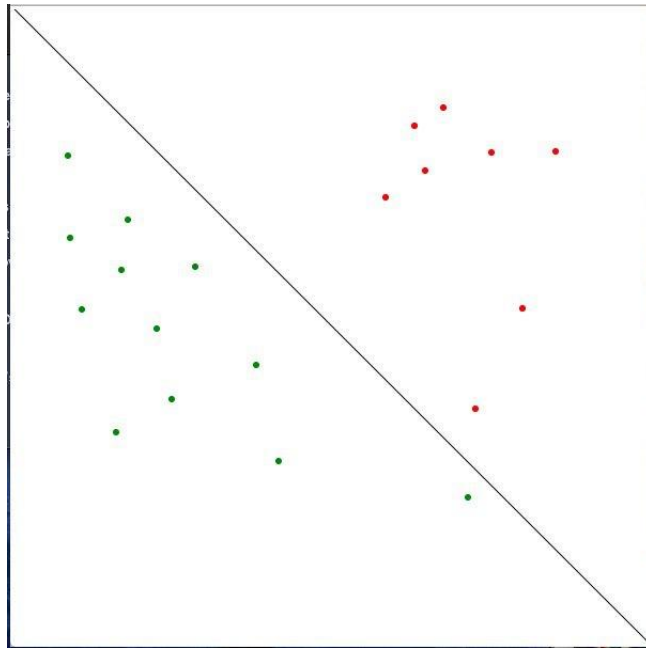
Тепер перейдемо безпосередньо до результатів роботи програми:

Очевидно точність розв'язку залежатиме від двох параметрів:

- 1) - кількість заданих точок
- 2) - кількість ітерацій методу оптимізації

Тут кількість точок дорівнює 20, кількість ітерацій - 1000. Тепер для прикладу візьмемо схожий за розташування набір точок, але зробимо лише 10 ітерацій.

Результат наступний:



Явно можемо бачити, що 10 ітерацій метода та 20 точок ще мало для того, аби роз'язок повністю нас задовільнив.

Варто зазначити, що для усіх прикладів $\theta^0 = [10, -10, 10]$, а початок координат - це є верхній лівий кут.

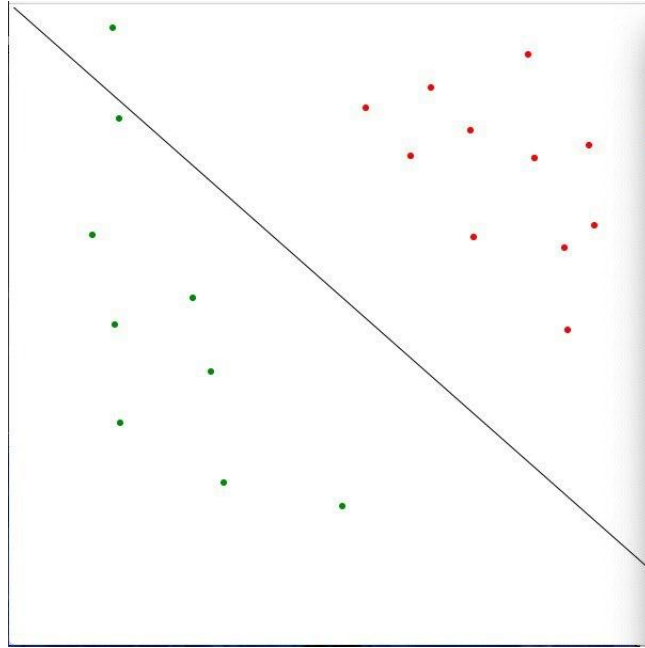
Отже ми побачили різницю між кількістю ітерацій методу, але постає інше питання - чи покращиться робота, якщо ми збільшимо кількість точок. Наприклад візьмемо 50 та 100 ітерацій.

Варто зазначити, що для цього прикладу ми тестуємо тільки метод градієнтного спуску, а метод лінійного каплінгу розглянемо трохи згодом для дещо ускладненої задачі.

Тепер давайте залишимо таку саму кількість ітерацій, але кількість початкових точок збільшимо до 50.

Результат наступний:

Тут ми вже ясно можемо сказати, що навіть з малою кількістю ітерацій метода градієнтного спуску, але з великою кількістю точок ми вже можемо говорити про непогану точність.



7.2 Приклад для тривимірного розділення точок

Тут ми розглянемо ту саму задачу, що і в попередньому розділі, але вона вже буде для тривимірного варіанту.

Тепер точки задав програмно, тобто їх місцезнаходження задано в кодї програми. Не обмежуючи загальності я поставив точки на діагоналі, тобто вони всі мають вигляд $[a, a, a]$, де $a \in [1, 20]$

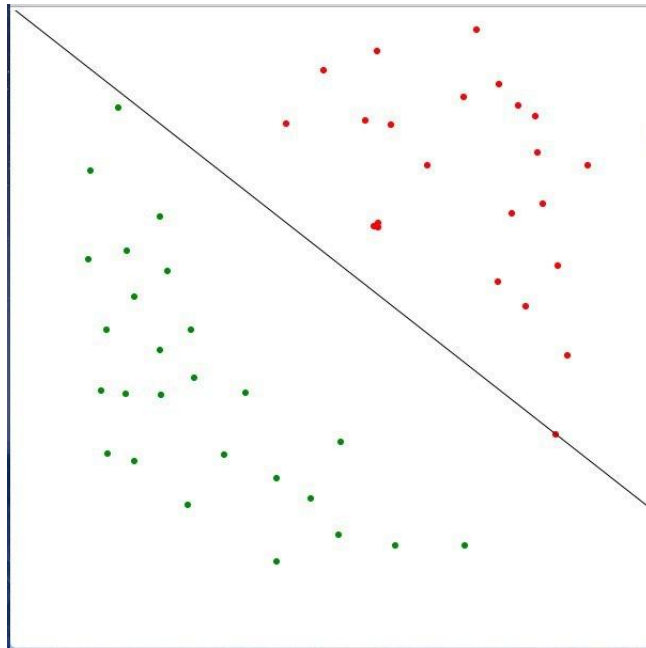
Маємо наступний результат:

Аби побудувати синій прямокутник я мав задати 4 точки, які відповідали б рівнянню площини для отриманої θ , тобто:

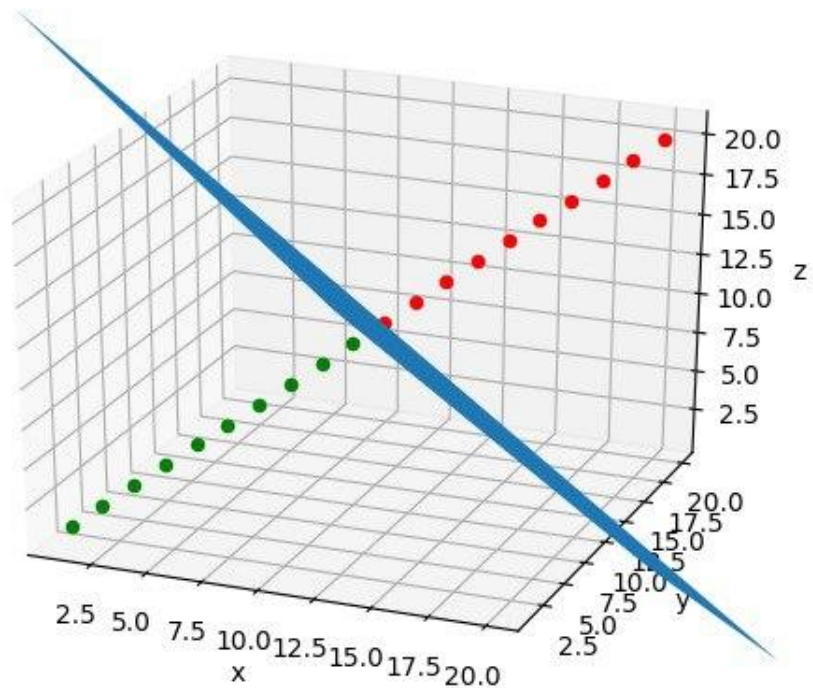
$$\theta_0 + \theta_1 x + \theta_2 y + \theta_3 z = 0$$

Сам алгоритм працював 1000 ітерацій і на виході я отримав

$$\theta = [300.67, -9.34, -9.34, -9.34]$$



. Потім я просто обираю довільні дві змінні та фіксував їх, а знаючи значення θ я мав змогу отримати значення третьої змінної, тобто точку, яка б належала шуканій площині. Зафіксувавши точки подалі одна від одної я і отримав таку площину. $\theta_0 = [10, 10, 10, 10]$. Знайдеі точки для прямокутника: $[0, 0, 32.3]$, $[0, 32.2, 0]$, $[20, 20, -7.7]$, $[32.3, 0, 0]$



8 Лінійна регресія

Тепер розглянемо інший приклад для наших методів - задачу лінійної регресії.

Лінійна регресія моделює залежну змінну як лінійну комбінацію незалежних змінних з доданим зсувом. Отже лінійна регресійна модель може бути записана у вигляді:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots \theta_n x_n$$

, де y - залежна змінна, x_1, x_2, \dots, x_n - незалежні змінні, θ_0 - зсув, $\theta_1, \theta_2, \dots, \theta_n$ - параметри моделі, що відповідають коефіцієнтам перед незалежними змінними. При налаштуванні моделі лінійної регресії необхідно визначити початкові значення параметрів $\theta_0, \theta_1, \theta_2, \dots, \theta_n$. Існує кілька підходів для генерації

початкових параметрів, зокрема випадковий вибір або використання нульових значень. У роботі розглянуто випадок випадкової генерації.

Основна задача в лінійній регресії - знайти оптимальні значення параметрів $\theta_0, \theta_1, \theta_2, \dots, \theta_n$, які мінімізують функцію втрат. Зазвичай, для лінійної регресії використовують середньоквадратичну помилку як функцію втрат, яка вимірює середнє квадратичне відхилення прогнозованих значень від реальних значень.

Оптимізація параметрів може бути виконана за допомогою методів градієнтного спуску або важкої кульки, які шукають локальний мінімум функції втрат. Ці методи поступово оновлюють значення параметрів, керуючись градієнтом функції втрат. В якості Функції втрат для лінійної регресії, візьмемо середньо квадратичну помилкою (Mean Squared Error, MSE). Ця функція вимірює різницю між прогнозованими значеннями моделі і фактичними значеннями залежної змінної. Загальний вигляд функції втрат:

$$MSE = \frac{1}{m} \sum (y_i - \hat{y}_i)^2,$$

де m - кількість спостережень (вибірка), y_i - фактичне значення залежної змінної для i -го спостереження, \hat{y}_i - прогнозоване значення залежної змінної для i -го спостереження. Також для прикладу розглянемо тривимірний випадок. Для нього усе буде аналогічно, просто збільшимо розмірність почтакових даних.

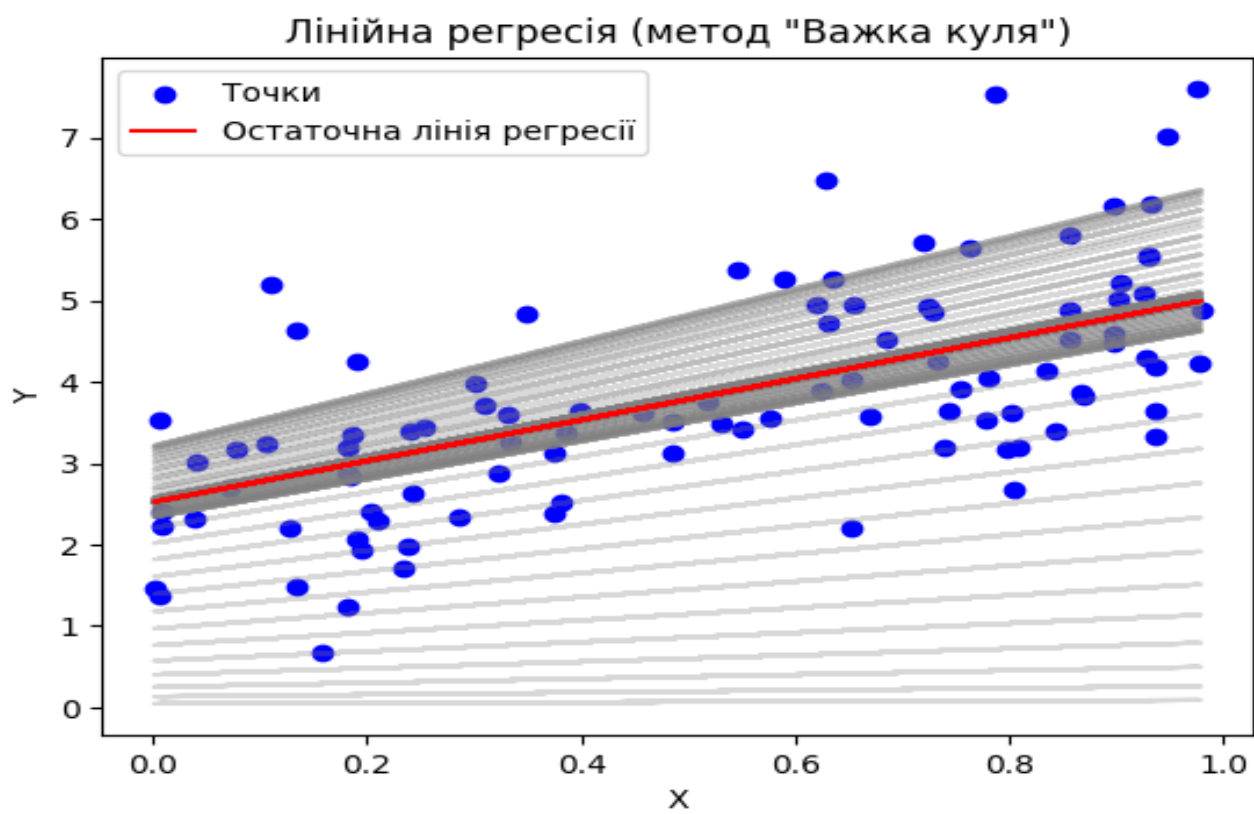
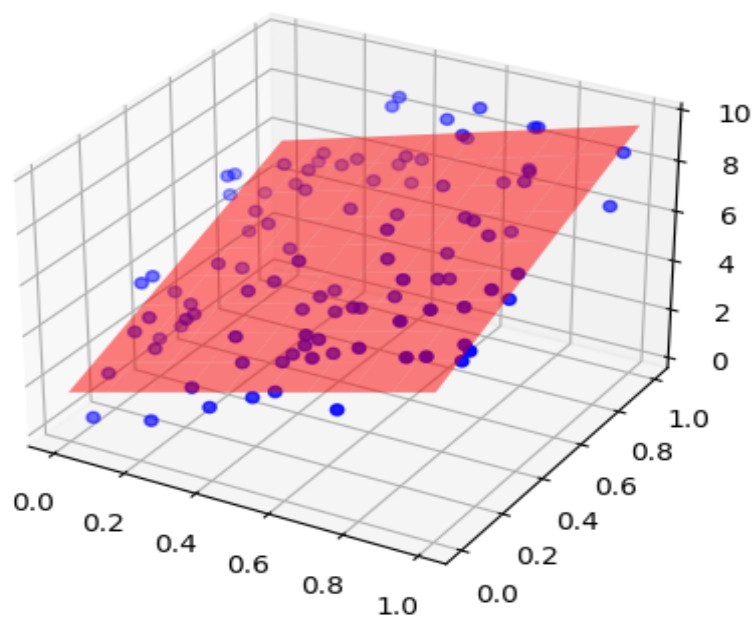


Рис. 4: Сірий - кроки регресії, червоний - остаточна відповідь



9 Висновок

У даній роботі були розглянуті теоретично та на практичних прикладах методи градієнтного спуску, лінійного каплінгу та важкої кульки. На практичних результатах, для різних прикладів, було отримано результати, з яких можна зробити висновки про швидкість та ефективність розглянутих методів і теоретичні результати збіглися із практичними результатами. Усі програми були написані використовуючи мову програмування Python3.0, графіки (2d, 3d) були отримані за допомогою бібліотеки matplotlib.

10 Джерела

- [1] Zeyuan Allen-Zhu, Lorenzo Orecchia, *Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent*, arXiv:1407.1537v5, 2016.
- [2] Yurii Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In Doklady AN SSSR (translated as Soviet Mathematics Doklady), volume 269, P 543–547, 1983.
- [3] Yurii Nesterov, *Smooth minimization of non-smooth functions*. *Mathematical Programming*, 103(1):127–152, December 2005.
- [4] John C. Duchi, *Introductory Lectures on Stochastic Optimization*, P. 40-50, 2010
- [5] Semenov V.V, *A Version of the Mirror descent Method to Solve Variational Inequalities*. *Cybernetics and Systems Analysis* 2017. Vol. 53. Issue 2. P. 234-243
- [6] <https://en.wikipedia.org/>, Logistic regression