Sprawozdanie z przeszukiwania przestrzeni – Aleksander Pizoń

Wstęp

Metoda gradientu prostego jest narzędziem optymalizacyjnym, wykorzystywanym do minimalizacji funkcji celu poprzez iteracyjne kierowanie się w kierunku przeciwnym do gradientu tej funkcji. Metoda charakteryzuje się prostotą implementacji, szybkością działania oraz możliwością skalowania jej to problemów wielu zmiennych. Do jej wad należą: brak gwarancji znalezienia minimum lokalnego, wrażliwość na dobór kroku uczenia oraz podatność na utykanie w punktach siodłowych dla funkcji wielu zmiennych.

Gradient funkcji wielu zmiennych to wektor zawierający pochodne cząstkowe tej funkcji po każdej zmiennej. Kierunek gradientu wskazuje największy wzrost funkcji, a kierunek przeciwny - największy spadek. Metoda gradientu prostego wykorzystuje ten fakt, poruszając się po przestrzeni parametrów funkcji celu w kierunku, który prowadzi do spadku wartości funkcji, aż do osiągnięcia minimum lokalnego lub globalnego.

W praktyce, algorytm ten wymaga ustalenia odpowiednich parametrów, takich jak parametr uczenia β, który decyduje o tym, jak duży krok wykonujemy w każdej iteracji, oraz kryterium stopu, które określa warunek zakończenia iteracji.

W tym zadaniu będziemy stosować metodę gradientu prostego w celu znalezienia minimum funkcji celu, poszukując optymalnego rozwiązania dla danego problemu.

Wzory metody gradientu prostego

Algorytm gradientu prostego opiera się na wzorze:

$$x_{k+1} = x_k - \nabla J(x_k) * \beta$$

Gdzie x_k to współrzędne obecnego punktu, $\nabla J(x_k)$ to wartość gradientu w punkcie x_k, β beta to parametr uczenia a x_{k+1} to wyliczone współrzędne następnego punktu.

Dla funkcji kwadratowej:

$$J = x^2$$

$$\nabla J = 2x$$

Dla funkcji Rastringa dla d = 2:

$$J = 20 + \sum_{i=1}^{2} [x_i^2 - 10\cos(2\pi x_i)]$$

$$\nabla J = \left[\frac{\partial J}{\partial x_1}, \frac{\partial J}{\partial x_2}\right]$$

$$\nabla J = \ [2x_1 + 20\pi \sin(2\pi x_1), 2x_2 + 20\pi \sin(2\pi x_2)\ ]$$

Dla funkcji Griewanka dla d = 2:

$$J = \sum_{i=1}^{2} \frac{x_i^2}{4000} - \prod_{i=1}^{2} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$$

$$\nabla J = \left[\frac{\partial J}{\partial x_1}, \frac{\partial J}{\partial x_2}\right]$$

$$\nabla J = \left[\frac{x_1}{2000} + \sin x_1 * \cos\left(\frac{x_2\sqrt{2}}{2}\right), \frac{x_2}{2000} + \sin x_2 * \cos\left(\frac{x_1\sqrt{2}}{2}\right)\right]$$

Opis algorytmu i eksperymentów

Napisany przeze mnie program wykonywany jest poprzez uruchomienie pliku main.py. Program polega na wykonaniu 3 eksperymentów dotyczących przeszukiwania przestrzeni parametrów. Argumenty potrzebne do wykonania eksperymentów znajdują się w pliku parameters.py. Implementacje funkcji wykorzystywanych w eksperymentach zawarte są w pliku functions.py. W pliku plots.py znajdują się funkcje odpowiedzialne za stworzenie i zapisanie wykresów pokazujących przebiegi eksperymentów, a w pliku save_results.py funkcje odpowiedzialne za zapisanie wyników eksperymentów do plików tekstowych.

Właściwy algorytm gradientu prostego jest zaimplementowany w pliku solver.py. Znajdują się w nim 2 klasy – Solver oraz Function. Aby móc zastosować metodę gradientu prostego do funkcji (konieczne jest, aby funkcja była jako obiekt obsługiwany przez bibliotekę sympy oraz aby zmienne były podane w postaci ‘x[i]’ gdzie i to indeks zmiennej liczony od 1) należy utworzyć obiekt klasy Solver podając jako argumenty przy tworzeniu odpowiednio: obiekt funkcji celu którą chcemy optymalizować (fun), ilość wymiarów optymalizowanej funkcji (dim), punkt początkowy (x0) oraz parametr uczenia (beta). Podczas inicjacji obiektu atrybutowi J zostaje przypisany obiekt klasy Function którego zadaniem jest przede wszystkim zwracanie wartości gradientu funkcji celu dla określonego punktu.

Najważniejszą metodą klasy Solver jest step_till_stop. Po jej wywołaniu wykonywane są kroki w kierunku ujemnego gradientu do momentu aż nie zostanie spełniony warunek stopu. W swoim programie umieściłem dwa warunki stopu: odległość progową trsh (gdy kolejne kroki są od siebie odległe o wartość mniejszą niż odległość progowa zostaje spełniony warunek stopu) oraz maksymalną liczbę kroków jaka może zostać wykonana przed zatrzymaniem się optymalizacji.

Eksperymenty

Eksperyment 1 był prowadzony dla funkcji kwadratowej dla różnych wartości parametru uczenia.

Eksperyment 2 był prowadzony dla funkcji Rastringa dla 10 punktów początkowych i przy 4 różnych wartościach parametru uczenia. Dane uzyskane podczas przeprowadzenia eksperymentu zamieszczono w tabeli 1. Na zamieszczonych pod tabelą wykresach.

Tabela 1.

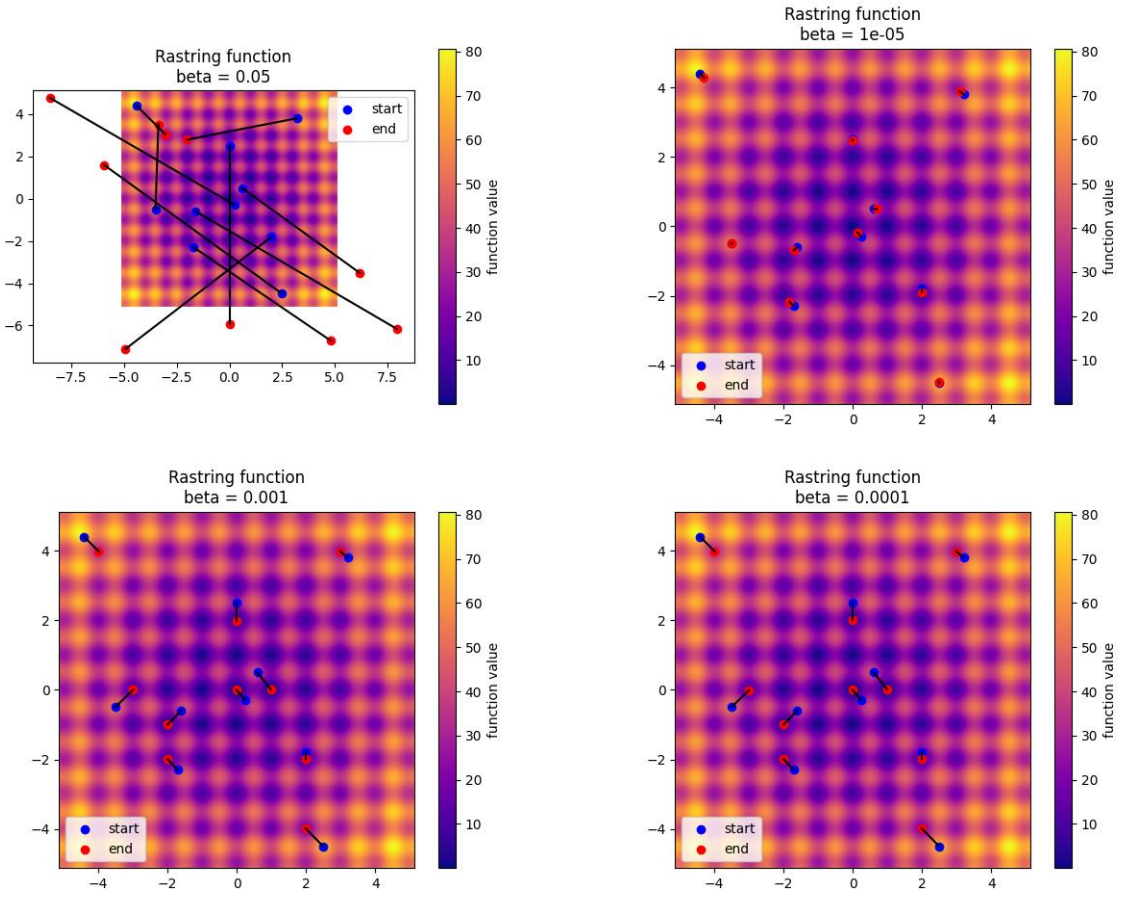
x1 pocz.	x2 pocz.	x1 końc.	x2 końc.	beta	kroki	LSD	MSE	L1
-4,4	4,4	-3,03777	3,037766	0,05	200	4,668117	18,46	6,076
-4,4	4,4	-3,97978	3,979784	0,001	200	0	31,68	7,96
-4,4	4,4	-3,98006	3,980062	0,0001	200	1,61E-05	31,68	7,96
-4,4	4,4	-4,28118	4,281185	1,00E-05	200	0,000992	36,66	8,562
-3,5	-0,5	-3,3633	3,503028	0,05	200	3,290076	23,58	6,866
-3,5	-0,5	-2,98486	-6,74E-42	0,001	200	4,44E-42	8,909	2,985
-3,5	-0,5	-2,98693	-0,01549	0,0001	200	0,000645	8,922	3,002
-3,5	-0,5	-3,4788	-0,49697	1,00E-05	200	0,000154	12,35	3,976
-1,7	-2,3	4,81477	-6,72797	0,05	200	3,522135	68,45	11,54
-1,7	-2,3	-1,98991	-1,98991	0,001	200	0	7,92	3,98
-1,7	-2,3	-1,98977	-1,99005	0,0001	200	8,05E-06	7,919	3,98
-1,7	-2,3	-1,81568	-2,16956	1,00E-05	200	0,000803	8,004	3,985
-1,6	-0,6	7,93136	-6,16079	0,05	200	4,173115	100,9	14,09
-1,6	-0,6	-1,98991	-0,99496	0,001	200	0	4,95	2,985
-1,6	-0,6	-1,98957	-0,99463	0,0001	200	1,96E-05	4,948	2,984
-1,6	-0,6	-1,6894	-0,69439	1,00E-05	200	0,000795	3,336	2,384
0	2,5	0	-5,94475	0,05	200	1,074158	35,34	5,945
0	2,5	0	1,989912	0,001	200	0	3,96	1,99
0	2,5	0	1,992817	0,0001	200	0,00012	3,971	1,993
0	2,5	0	2,484852	1,00E-05	200	0,000109	6,174	2,485
0,25	-0,3	-8,48877	4,756596	0,05	200	2,001526	94,68	13,25
0,25	-0,3	4,91E-45	-7,49E-45	0,001	200	5,89E-45	8E-89	1E-44
0,25	-0,3	9,81E-05	-0,00014	0,0001	200	6,92E-06	3E-08	2E-04
0,25	-0,3	0,13497	-0,1769	1,00E-05	200	0,00074	0,05	0,312
0,6	0,5	6,16079	-3,50303	0,05	200	4,93022	50,23	9,664
0,6	0,5	0,99496	6,74E-42	0,001	200	4,44E-42	0,99	0,995
0,6	0,5	0,99463	0,015491	0,0001	200	0,000639	0,99	1,01

0,6	0,5	0,69439	0,49697	1,00E-05	200	0,000576	0,729	1,191
2	-1,8	-4,94319	-7,12645	0,05	200	3,099965	75,22	12,07
2	-1,8	1,98991	-1,98991	0,001	200	0	7,92	3,98
2	-1,8	1,98992	-1,98984	0,0001	200	2,91E-06	7,919	3,98
2	-1,8	1,99447	-1,8929	1,00E-05	200	0,000355	7,561	3,887
2,5	-4,5	-5,94475	1,597826	0,05	200	3,783673	37,89	7,543
2,5	-4,5	1,98991	-3,97978	0,001	200	0	19,8	5,97
2,5	-4,5	1,99282	-3,98142	0,0001	200	0,000137	19,82	5,974
2,5	-4,5	2,48485	-4,47275	1,00E-05	200	0,000224	26,18	6,958
3,2	3,8	-2,01563	2,785645	0,05	200	3,484842	11,82	4,801
3,2	3,8	2,98486	3,979784	0,001	200	0	24,75	6,965
3,2	3,8	2,98493	3,979712	0,0001	200	4,28E-06	24,75	6,965
3,2	3,8	3,09168	3,88666	1,00E-05	200	0,000525	24,66	6,978

LSD – last steps distance– odległość między położeniami punktów podczas ostatnich dwóch kroków

MSE – mean squared error

L1 – pierwsza norma wektora odległości między wartością oczekiwaną a wartością uzyskaną



Eksperyment 3 był prowadzony dla funkcji Griewanka dla 10 punktów początkowych i przy 4 różnych wartościach parametru uczenia. Dane uzyskane podczas przeprowadzenia eksperymentu zamieszczono w tabeli 2. Na zamieszczonych pod tabelą wykresach.

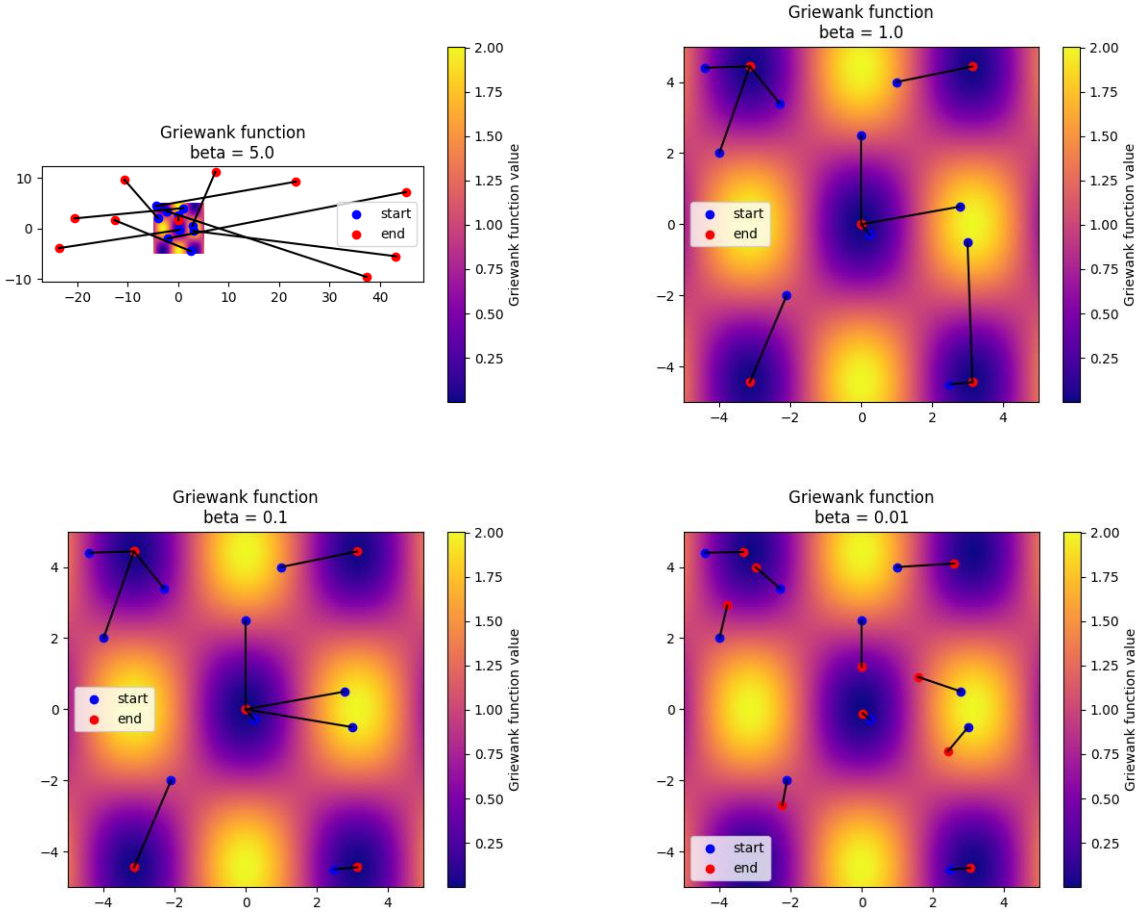
Tabela 2.

x1 pocz.	x2 pocz.	x1 końc.	x2 końc.	beta	kroki	LSD	MSE	L1
0,25	-0,3	-23,56200881	-3,907448925	5	200	3,00E+00	570,436416	27,46946
0,25	-0,3	0	-1,59132E-61	1	200	1,59E-61	2,532E-122	1,59E-61
0,25	-0,3	1,80105E-10	-1,05403E-05	0,1	200	5,55E-07	1,111E-10	1,05E-05
0,25	-0,3	0,034312278	-0,111209241	0,01	200	6,57E-04	0,01354483	0,145522
2,8	0,5	7,36696118	11,18263426	5	200	1,76E+00	179,323426	18,5496
2,8	0,5	0	1,09585E-59	1	200	1,10E-59	1,201E-118	1,1E-59
2,8	0,5	1,98141E-08	0,000154881	0,1	200	8,16E-06	2,3988E-08	0,000155
2,8	0,5	1,601349082	0,906944335	0,01	200	8,02E-03	3,38686691	2,508293
-4,4	4,4	23,29505414	9,224678294	5	200	3,17E+00	627,754237	32,51973
-4,4	4,4	-3,140022634	4,438444481	1	200	0,00E+00	29,5595316	7,578467
-4,4	4,4	-3,140022635	4,438442745	0,1	200	9,14E-08	29,5595162	7,578465
-4,4	4,4	-3,334885998	4,420575606	0,01	200	1,96E-03	30,6629533	7,755462
0	2,5	0	1,603889641	5	200	3,21E+00	2,57246198	1,60389
0	2,5	0	2,65016E-60	1	200	2,66E-60	7,023E-120	2,65E-60
0	2,5	0	0,000122455	0,1	200	6,45E-06	1,4995E-08	0,000122
0	2,5	0	1,190415203	0,01	200	5,30E-03	1,41708836	1,190415
-2,3	3,4	37,25526044	-9,593108024	5	200	2,80E+00	1479,98215	46,84837
-2,3	3,4	-3,140022634	4,438444481	1	200	0,00E+00	29,5595316	7,578467
-2,3	3,4	-3,140022633	4,43840158	0,1	200	2,26E-06	29,5591507	7,578424
-2,3	3,4	-2,982863518	3,996434204	0,01	200	2,63E-03	24,8689611	6,979298
2,5	-4,5	-12,57661339	1,601653425	5	200	3,21E+00	160,736498	14,17827
2,5	-4,5	3,140022634	-4,438444481	1	200	0,00E+00	29,5595316	7,578467
2,5	-4,5	3,140022633	-4,438446732	0,1	200	1,19E-07	29,5595515	7,578469
2,5	-4,5	3,051220185	-4,462106139	0,01	200	9,04E-04	29,2203358	7,513326
-2,1	-2	45,08658706	7,171951971	5	200	5,57E-01	2084,23723	52,25854
-2,1	-2	-3,140022634	-4,438444481	1	200	0,00E+00	29,5595316	7,578467
-2,1	-2	-3,140022613	-4,438194973	0,1	200	1,31E-05	29,5573166	7,578218
-2,1	-2	-2,232286599	-2,699138998	0,01	200	4,83E-03	12,2684548	4,931426
3	-0,5	43,12749822	-5,527135003	5	200	4,22E+00	1890,53032	48,65463
3	-0,5	3,140022634	-4,438444481	1	200	0,00E+00	29,5595316	7,578467
3	-0,5	6,39791E-07	-0,001259251	0,1	200	6,63E-05	1,5857E-06	0,00126
3	-0,5	2,424273262	-1,187105958	0,01	200	5,92E-03	7,2863214	3,611379
1	4	-20,47281014	1,967671221	5	200	1,86E+00	423,007685	22,44048
1	4	3,140022634	4,438444481	1	200	0,00E+00	29,5595316	7,578467
1	4	3,14002263	4,438408777	0,1	200	1,88E-06	29,5592146	7,578431
1	4	2,607512576	4,096431681	0,01	200	5,18E-03	23,5798743	6,703944
-4	2	-10,62562449	9,583331305	5	200	4,12E+00	204,744135	20,20896
-4	2	-3,140022634	4,438444481	1	200	0,00E+00	29,5595316	7,578467
-4	2	-3,140022641	4,43826334	0,1	200	9,54E-06	29,5579237	7,578286
-4	2	-3,780256055	2,927803959	0,01	200	5,74E-03	22,8623719	6,70806

LSD – last steps distance– odległość między położeniami punktów podczas ostatnich dwóch kroków

MSE – mean squared error

L1 – pierwsza norma wektora odległości między wartością oczekiwaną a wartością uzyskaną



Wnioski

Pierwszym nasuwającym się wnioskiem jest brak gwarancji na to, że zaimplementowany algorytm gradientu prostego odnajdzie minimum globalne funkcji celu. Jak wynika z danych dla funkcji Rastringa, jedynie dla punktu początkowego x1 = 0.25, x2 = -0.3 algorytm odnajduje minimum globalne. Dla wszystkim pozostałych 9 punktów znajdujących się dalej od minimum globalnego algorytm zatrzymuje się w najbliższym minimum lokalnym. Podczas optymalizacji funkcji Griewanka zachodzi podobne zjawisko. Jest to spowodowane tym, że optymalizowane funkcje nie są dostatecznie gładkie oraz posiadają wiele minimów lokalnych. Sam algorytm z kolei, gdy znajdzie się w minimum lokalnym nie jest w stanie z niego wyjść, ponieważ w każdym ekstremum wartość gradientu wynosi zero.

Kolejnym wnioskiem jest znaczący wpływ parametru uczenia β na skuteczność optymalizacji przez metodę gradientu prostego. W przypadku optymalizacji funkcji celu za duża wartość parametru uczenia (w przypadku funkcji Rastringa jest to już beta=0.5, a dla funkcji Griewanka beta=5) sprawia, że algorytm nie jest w stanie odnaleźć żadnego minimum. Algorytm w kolejnych krokach przeskakuje minimum, do którego prowadzi wzięty z minusem gradient i znajduje się w nowym punkcie, w którym gradient prowadzi już do innego ekstremum. Z kolei zbyt mała wartość parametru powoduje, że algorytm zbyt wolno zbliża się do minimum co pociąga za sobą konieczność zadania programowi nieoptymalnie dużej ilości kroków.

Innym ważnym do zaobserwowania faktem jest zależność optymalnego parametru uczenia od minimalizowanej funkcji celu. Dla funkcji Rastringa wartość parametru wynosząca 0.001 sprawia, że algorytm bez trudu odnajduje minimum podczas gdy dla funkcji Griewanka już wartość 0.01 jest zbyt mała, aby program w optymalnej liczbie kroków odnalazł minimum. Z kolei optymalna wartość beta dla funkcji Griewanka wynosząca 1.0 jest już znacząco za duża dla funkcji Rastringa i uniemożliwia znalezienia jej minimum.

Podsumowując, metoda gradientu prostego nie gwarantuje odnalezienia minimum globalnego funkcji celu a sama skuteczność optymalizacji bardzo mocno zależy od wybranej wartości parametru uczenia. Konieczność doboru wartości tego parametru osobno dla każdej optymalizowanej funkcji jest spowodowana różną gładkością funkcji. Dla funkcji Rastringa wartość beta musi być odpowiednio mała, ponieważ powierzchnia funkcji jest mocno pofalowana i ma wiele minimów. Dla funkcji Griewanka ten parametr musi być odpowiednio większy.