MDS 6106 Assignment 4

Guyuan Xu   224040074

A4.1 (Implementing the Gradient Method)

We want to minimize the objective function

$$\min_{x\in\mathbb{R}^2} f(x) = \frac{1}{2}x_1^4 - x_1^3 - x_1^2 + x_1^2x_2^2 + \frac{1}{2}x_2^4 - x_2^2$$

by gradient descent methods with different initial points and stepsize strategies, as presented in the following.

Initial points:

$$\chi^0 := \left\{ \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix}, \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \begin{bmatrix} -\frac{1}{4} \\ -\frac{1}{2} \end{bmatrix}, \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}, \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix} \right\}$$

stationary points of $f(x)$:

$$\chi^* := \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} -\frac{1}{2} \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}$$

1. Backtracking

Back Tracking line search: choose the largest $\alpha_k \in \{\sigma^k : k = 0, 1, ...\}$ that satisfies Armijo condition $f(x_k + \alpha_k d_k) - f(x_k) \leq \gamma\alpha_k \nabla f(x_k)^T d_k$ with $(\sigma, \gamma) = (0.5, 0.1)$.

Performance (in terms of iteration) of Backtracking Line Search stepsize strategy are shown below:

x_0	iteration	limit point x^*
(-0.50,1.00)	13	(2.00,-0.00)
(-0.50,0.50)	325	(-0.00,1.00)
(-0.25,-0.50)	467	(-0.00,-1.00)
(0.50,-0.50)	12	(2.00,0.00)
(0.50,1.00)	10	(2.00,-0.00)

Backtracking Line Search

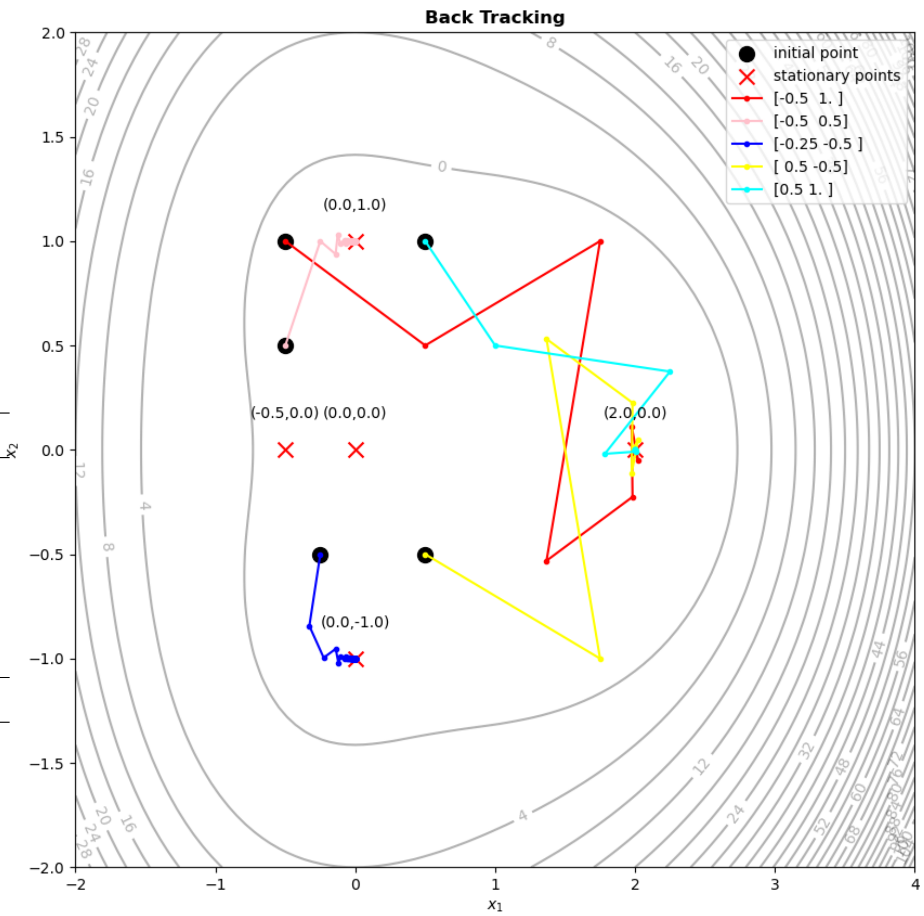


Figure 1: Backtracking Line Search

2. Exact Line Search

Exact Line Search: aim at choosing the stepsize α_k that

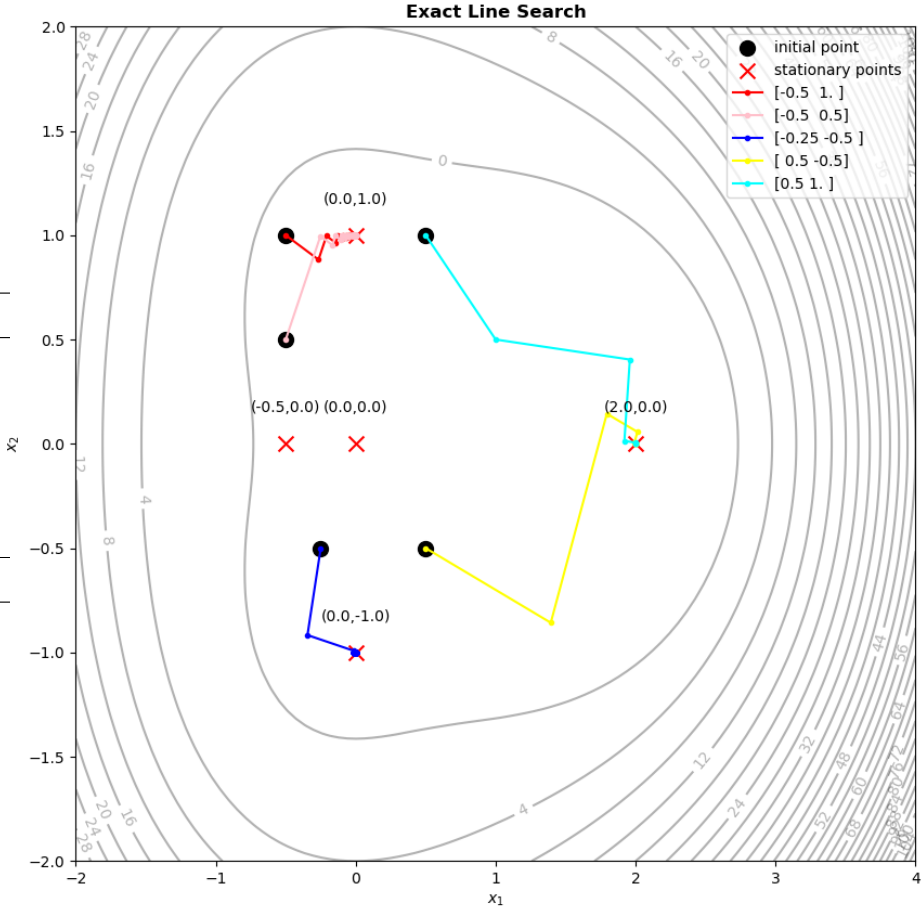
$$\alpha_k = \underset{\alpha_k \geq 0}{argmin}\ f(x_k + \alpha_k d_k)$$

Performance of Exact line search stepsize strategy are shown below:

x_0	iteration	limit point x^*
(-0.50,1.00)	295	(-0.00,1.00)
(-0.50,0.50)	296	(-0.00,1.00)
(-0.25,-0.50)	375	(-0.00,-1.00)
(0.50,-0.50)	9	(2.00,0.00)
(0.50,1.00)	6	(2.00,0.00)

Exact Line Search

PS: the paths of 2 consecutive steps are not perpendicular because we constraint $\alpha_k \leq 1$ (because we search α in $[0, a]$, where $a = 1$) insetead of not setting any constraint to them.



3. Diminishing Stepsize

Diminishing Stepsize: we simply set

$$\alpha_k = \frac{1}{\sqrt{k+2}}$$

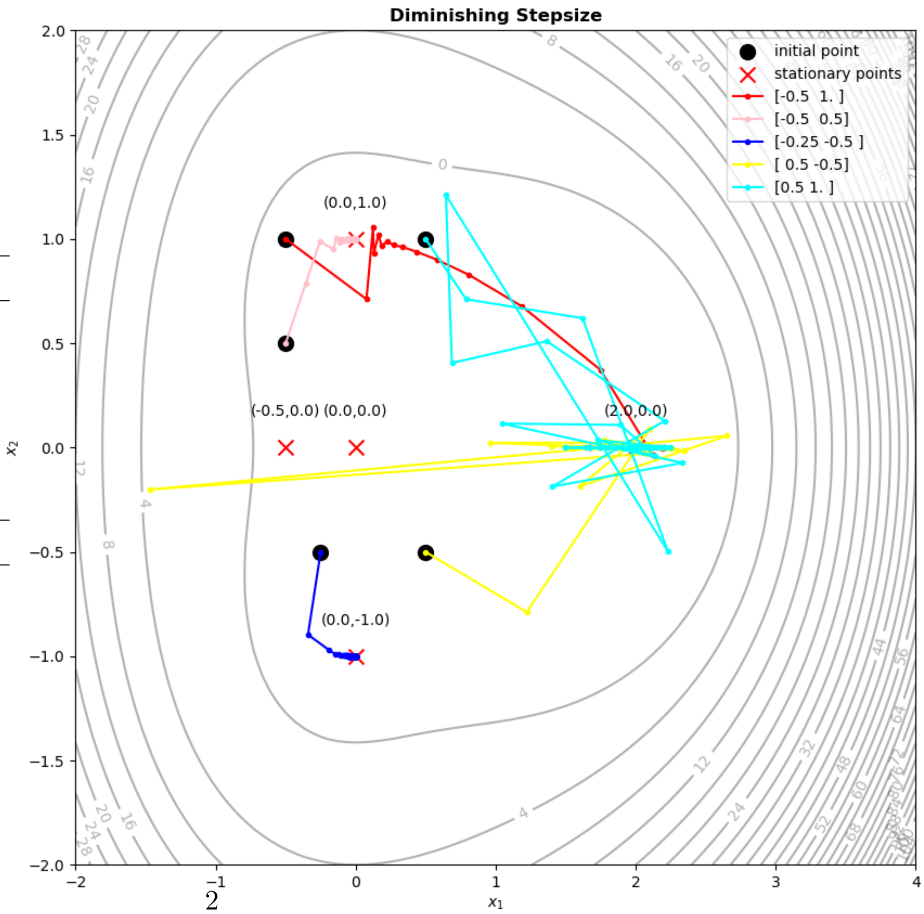
where k is the round of iteration.
Performance of diminishing stepsize strategy are shown below:

x_0	iteration	limit point x^*
(-0.50,1.00)	47	(2.00,0.00)
(-0.50,0.50)	8523	(-0.00,1.00)
(-0.25,-0.50)	8501	(-0.00,-1.00)
(0.50,-0.50)	47	(2.00,-0.00)
(0.50,1.00)	47	(2.00,0.00)

Diminishing Stepsize

PS: Since HW sheet has no requirement on k, we follow the common sense that k starts from 1, so the first stepsize is

$$\alpha_1 = \frac{1}{\sqrt{1+2}} = \frac{1}{\sqrt{3}}$$

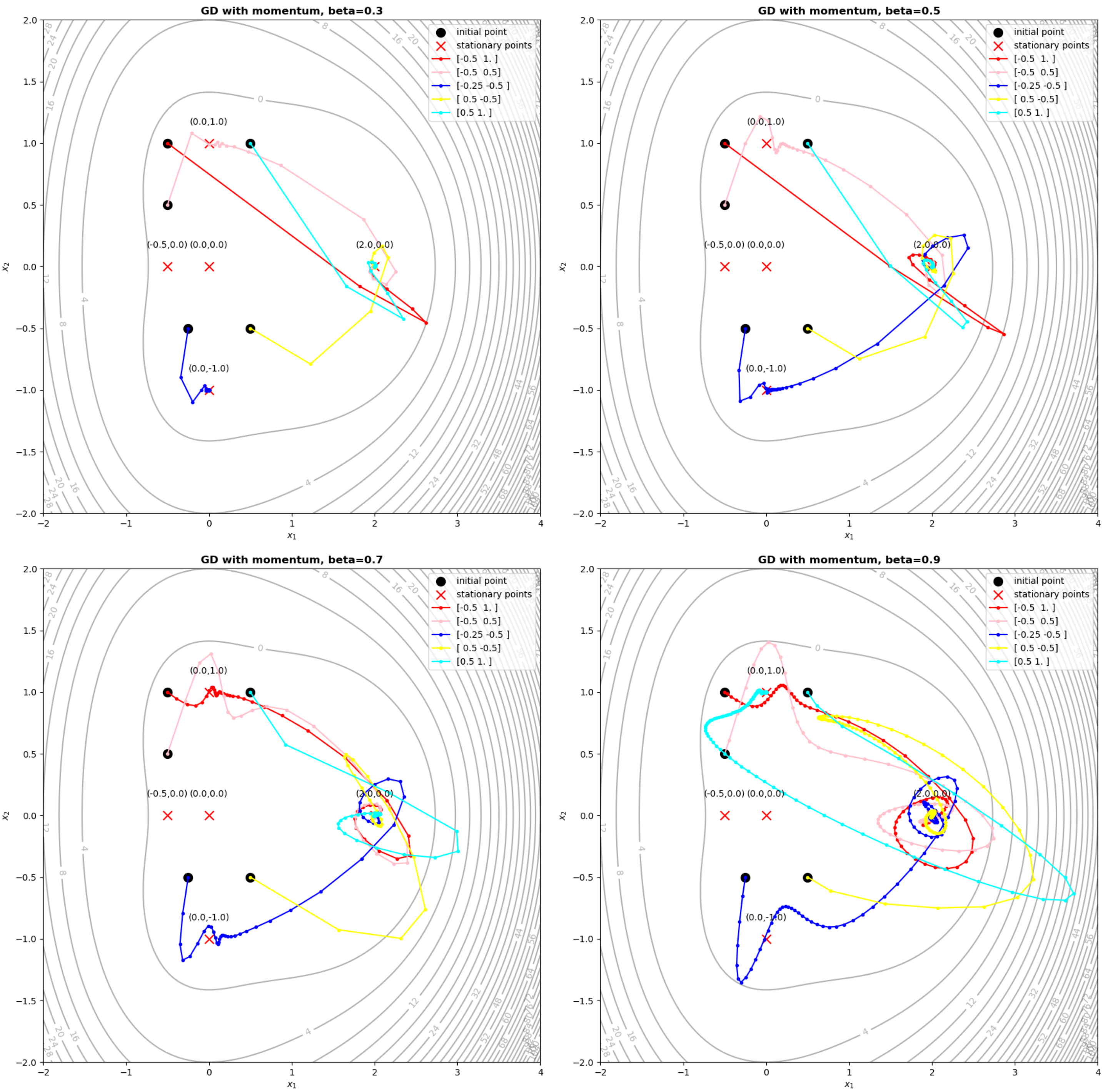


Then we can conclude the performance of different stepsize strategies as the following table:

Methods	x_0	iteration	limit point x^*	Global Minimum?
Back Tracking	(-0.50,1.00)	13	(2.00,-0.00)	yes
	(-0.50,0.50)	325	(-0.00,1.00)	no
	(-0.25,-0.50)	467	(-0.00,-1.00)	no
	(0.50,-0.50)	12	(2.00,0.00)	yes
	(0.50,1.00)	10	(2.00,-0.00)	yes
Exact Line Search	(-0.50,1.00)	295	(-0.00,1.00)	no
	(-0.50,0.50)	296	(-0.00,1.00)	no
	(-0.25,-0.50)	375	(-0.00,-1.00)	no
	(0.50,-0.50)	9	(2.00,0.00)	yes
	(0.50,1.00)	6	(2.00,0.00)	yes
Diminishing Stepsize	(-0.50,1.00)	47	(2.00,0.00)	yes
	(-0.50,0.50)	8523	(-0.00,1.00)	no
	(-0.25,-0.50)	8501	(-0.00,-1.00)	no
	(0.50,-0.50)	47	(2.00,-0.00)	yes
	(0.50,1.00)	47	(2.00,0.00)	yes

A4.2 (Inertial Gradient Method)

The convergence trace of gradient method with momentum of $\beta \in \{0.3, 0.5, 0.7, 0.9\}$ and different initial points are shown below respectively:



Performance Analysis

by comparing the average iteration numbers of GD with momentum and GD with different stepsize strategies (discussed in part A4.1):

	Stepsize strategies			GD with momentum			
	Backtrack	Exact LineSearch	Diminishing	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 0.9$
Average Iteration	164.5	196.2	3433.0	104.8	51.8	90.2	1250.0
Probablity to Global Min	0.6	0.4	0.6	0.8	1.0	1.0	0.8

It is easy to see that **the average iteration (averaging across 5 different initial points) it takes to converge to some limit point is noticeably larger using stepsize strategies than GD with momentum in general.**

And we also notice GD with momentum converger faster when β change from 0.3 to 0.5, then converger slower when β goes from 0.5 to 0.9, how to interpretate this phenomenon? we can first look at the detail of how β affect convergence:

β	x_0	iteration	limit point x^*	Global Minimum?
$\beta = 0.3$	(-0.50,1.00)	26	(2.00,0.00)	yes
	(-0.50,0.50)	33	(2.00,-0.00)	yes
	(-0.25,-0.50)	417	(-0.00,-1.00)	no
	(0.50,-0.50)	24	(2.00,0.00)	yes
	(0.50,1.00)	24	(2.00,0.00)	yes
$\beta = 0.5$	(-0.50,1.00)	39	(2.00,0.00)	yes
	(-0.50,0.50)	56	(2.00,0.00)	yes
	(-0.25,-0.50)	88	(2.00,0.00)	yes
	(0.50,-0.50)	37	(2.00,-0.00)	yes
	(0.50,1.00)	39	(2.00,0.00)	yes
$\beta = 0.7$	(-0.50,1.00)	105	(2.00,0.00)	yes
	(-0.50,0.50)	88	(2.00,0.00)	yes
	(-0.25,-0.50)	102	(2.00,-0.00)	yes
	(0.50,-0.50)	77	(2.00,-0.00)	yes
	(0.50,1.00)	79	(2.00,-0.00)	yes
$\beta = 0.9$	(-0.50,1.00)	268	(2.00,0.00)	yes
	(-0.50,0.50)	258	(2.00,0.00)	yes
	(-0.25,-0.50)	276	(2.00,0.00)	yes
	(0.50,-0.50)	297	(2.00,-0.00)	yes
	(0.50,1.00)	5151	(-0.00,1.00)	no

Notice when $\beta = 0.3$ and starting from the initial point $(-0.25, -0.5)$, GD converge to $(0, -1)$ while other initial points all converge to $(2, 0)$, this is because the first several stepsize of GD heppen to be too large from this initial point and **pushing the trace to an area not ideal for fast convergence**, and by coincidence the trace finally converge to a different limit point from the other initial points, and the iterations it takes to converge go very high, this should **be treated as an anomoly.** Same with $\beta = 0.9$ initial point $(0.5, 1.0)$: this very initial condition just happen to be not ideal for fast convergence.
After dropping this anomoly we can get an averge iteration number of 26.75 for $\beta = 0.3$, and the trend becomes obvious: β smaller, converge faster.

Another observation is that **GD with momentum is more likely to converge to the global minimum** $(2, 0)$ than GD with stepsize strategies, this is also easy to explain: with "momentum" (brought by the momentum term $\beta(x_k - k_{k-1})$), the trace is more likely to "escape" from the local minimum.