Homework 2

Matej Kalc

Exercise 5

For function 1 Newton got the minimum in just 2 steps with both initialization. In 0.1 the majority of methods achieved the minimum of f_1. For function 2 the first initialization gave better results sooner. For this function the best method was Newton. When checking the execution time the worst performing method was Adagrad GD, which probably needs more time to adapt the learning rate. For function 3 again the first initialization gave better results sooner then the second initialization. The best method was BFGS for this function. BFGS is the only method which reached the optimum minimum. For function 3 Newton method failed. Below are shown all the results for each function, option and method.

```
Function 1
n =  2
GD [-0.018 -0.02   0.001] -0.035
GD [ 0.687  0.951 -0.105] 6.041
PGD [-0.025 -0.03   0.002] -0.05
PGD [ 0.599  0.898 -0.141] 4.996
Nesterov GD [-0.025 -0.03   0.002] -0.05
Nesterov GD [ 0.6    0.898 -0.141] 5.0
Adagrad GD [-0.016 -0.017  0.01 ] -0.032
Adagrad GD [ 0.983  0.983 -0.017] 10.563
Newton [-0.167 -0.229  0.167] -0.198
Newton [-0.167 -0.229  0.167] -0.198
BFGS [-0.111 -0.111  0.    ] -0.111
BFGS [ 0.43  0.97 -0.18] 5.029
n =  5
GD [-0.037 -0.048  0.005] -0.073
GD [ 0.491  0.783 -0.19 ] 3.673
PGD [-0.043 -0.057  0.007] -0.083
PGD [ 0.448  0.723 -0.209] 3.153
Nesterov GD [-0.043 -0.057  0.007] -0.083
Nesterov GD [ 0.448  0.723 -0.209] 3.155
Adagrad GD [-0.03  -0.032  0.024] -0.056
Adagrad GD [ 0.968  0.968 -0.032] 10.185
Newton [-0.167 -0.229  0.167] -0.198
Newton [-0.167 -0.229  0.167] -0.198
BFGS [-0.122 -0.196  0.025] -0.156
BFGS [ 0.064  0.077 -0.459] 0.634
n =  10
GD [-0.061 -0.085  0.016] -0.113
GD [ 0.332  0.529 -0.25 ] 1.854
PGD [-0.065 -0.092  0.019] -0.119
PGD [ 0.307  0.483 -0.256] 1.61
Nesterov GD [-0.065 -0.091  0.019] -0.119
Nesterov GD [ 0.307  0.483 -0.256] 1.612
Adagrad GD [-0.044 -0.049  0.039] -0.081
Adagrad GD [ 0.95   0.95  -0.049] 9.756
Newton [-0.167 -0.229  0.167] -0.198
Newton [-0.167 -0.229  0.167] -0.198
BFGS [-0.114 -0.186  0.026] -0.157
BFGS [ 0.104  0.068 -0.447] 0.601
n =  100
GD [-0.16  -0.221  0.154] -0.198
GD [-0.141 -0.198  0.117] -0.192
PGD [-0.16  -0.222  0.155] -0.198
PGD [-0.143 -0.2    0.12 ] -0.193
Nesterov GD [-0.16  -0.222  0.155] -0.198
Nesterov GD [-0.143 -0.2    0.12 ] -0.193
Adagrad GD [-0.121 -0.158  0.119] -0.18
Adagrad GD [ 0.822  0.817 -0.171] 6.97
Newton [-0.167 -0.229  0.167] -0.198
Newton [-0.167 -0.229  0.167] -0.198
BFGS [-0.167 -0.229  0.167] -0.198
BFGS [-0.167 -0.229  0.167] -0.198
time max =  0.1
GD [-0.167 -0.229  0.167] -0.198
GD [-0.167 -0.229  0.167] -0.198
PGD [-0.167 -0.229  0.167] -0.198
PGD [-0.167 -0.229  0.167] -0.198
Nesterov GD [-0.167 -0.229  0.167] -0.198
Nesterov GD [-0.167 -0.229  0.167] -0.198
Adagrad GD [-0.11  -0.142  0.109] -0.171
Adagrad GD [ 0.85   0.846 -0.146] 7.523
Newton [-0.167 -0.229  0.167] -0.198
Newton [-0.167 -0.229  0.167] -0.198
BFGS [-0.167 -0.229  0.167] -0.198
BFGS [-0.167 -0.229  0.167] -0.198
time max =  1
GD [-0.167 -0.229  0.167] -0.198
GD [-0.167 -0.229  0.167] -0.198
PGD [-0.167 -0.229  0.167] -0.198
PGD [-0.167 -0.229  0.167] -0.198
Nesterov GD [-0.167 -0.229  0.167] -0.198
Nesterov GD [-0.167 -0.229  0.167] -0.198
Adagrad GD [-0.154 -0.209  0.154] -0.197
Adagrad GD [ 0.714  0.702 -0.267] 5.074
```

```
Newton [-0.167 -0.229  0.167] -0.198
Newton [-0.167 -0.229  0.167] -0.198
BFGS [-0.167 -0.229  0.167] -0.198
BFGS [-0.167 -0.229  0.167] -0.198
time max =  2
GD [-0.167 -0.229  0.167] -0.198
GD [-0.167 -0.229  0.167] -0.198
PGD [-0.167 -0.229  0.167] -0.198
PGD [-0.167 -0.229  0.167] -0.198
Nesterov GD [-0.167 -0.229  0.167] -0.198
Nesterov GD [-0.167 -0.229  0.167] -0.198
Adagrad GD [-0.162 -0.221  0.162] -0.198
Adagrad GD [ 0.658  0.642 -0.311] 4.259
Newton [-0.167 -0.229  0.167] -0.198
Newton [-0.167 -0.229  0.167] -0.198
BFGS [-0.167 -0.229  0.167] -0.198
BFGS [-0.167 -0.229  0.167] -0.198
Function 2
n =  2
GD [1.065 1.125 1.255] 0.042
GD [-1.023  1.133  1.217] 5.291
PGD [1.061 1.122 1.257] 0.02
PGD [-1.054  1.085  1.23 ] 4.563
Nesterov GD [1.061 1.122 1.257] 0.02
Nesterov GD [-1.054  1.085  1.23 ] 4.574
Adagrad GD [1.198 1.198 1.202] 11.211
Adagrad GD [-1.002  1.198  1.202] 13.33
Newton [1.058 1.114 1.229] 0.035
Newton [-0.891  0.745  0.327] 9.062
BFGS [1.079 1.129 1.25 ] 0.214
BFGS [-1.044  1.111  1.228] 4.24
n =  5
GD [1.059 1.121 1.258] 0.018
GD [-1.049  1.097  1.223] 4.243
PGD [1.059 1.121 1.257] 0.018
PGD [-1.044  1.106  1.219] 4.217
Nesterov GD [1.059 1.121 1.257] 0.018
Nesterov GD [-1.044  1.106  1.219] 4.218
Adagrad GD [1.197 1.197 1.203] 10.873
Adagrad GD [-1.003  1.197  1.203] 12.926
Newton [1.003 1.006 1.011] 0.0
Newton [-0.172 -0.099  0.005] 4.236
BFGS [1.061 1.119 1.258] 0.026
BFGS [-1.046  1.106  1.229] 4.215
n =  10
GD [1.059 1.121 1.257] 0.018
GD [-1.045  1.102  1.217] 4.203
PGD [1.059 1.121 1.257] 0.018
PGD [-1.045  1.101  1.217] 4.202
Nesterov GD [1.059 1.121 1.257] 0.018
Nesterov GD [-1.045  1.101  1.217] 4.202
Adagrad GD [1.195 1.195 1.205] 10.489
Adagrad GD [-1.005  1.195  1.205] 12.467
Newton [1. 1. 1.] 0.0
Newton [-2.533  5.771 30.021] 1156.428
BFGS [1.059 1.121 1.257] 0.018
BFGS [-1.048  1.106  1.226] 4.212
n =  100
GD [1.057 1.117 1.248] 0.017
GD [-1.029  1.069  1.147] 4.134
PGD [1.057 1.117 1.248] 0.017
PGD [-1.029  1.069  1.146] 4.133
Nesterov GD [1.057 1.117 1.248] 0.017
Nesterov GD [-1.029  1.069  1.146] 4.133
Adagrad GD [1.182 1.183 1.218] 7.939
Adagrad GD [-1.018  1.182  1.218] 9.525
Newton [1. 1. 1.] 0.0
Newton [1. 1. 1.] 0.0
BFGS [1. 1. 1.] 0.0
BFGS [0.617 0.382 0.147] 0.529
time max =  0.1
GD [1.026 1.052 1.108] 0.003
GD [0.305 0.096 0.008] 1.302
PGD [1.003 1.006 1.012] 0.0
PGD [0.98  0.961 0.923] 0.002
Nesterov GD [1.005 1.011 1.021] 0.0
Nesterov GD [0.961 0.924 0.854] 0.007
Adagrad GD [1.183 1.184 1.216] 8.183
Adagrad GD [-1.016  1.184  1.216] 9.796
Newton [1. 1. 1.] 0.0
Newton [1. 1. 1.] 0.0
BFGS [1. 1. 1.] 0.0
BFGS [1. 1. 1.] 0.0
time max =  1
GD [1. 1. 1.] 0.0
GD [1. 1. 1.] 0.0
PGD [1. 1. 1.] 0.0
PGD [1. 1. 1.] 0.0
Nesterov GD [1. 1. 1.] 0.0
Nesterov GD [1. 1. 1.] 0.0
Adagrad GD [1.169 1.173 1.229] 5.979
Adagrad GD [-1.026  1.173  1.227] 7.832
Newton [1. 1. 1.] 0.0
Newton [1. 1. 1.] 0.0
BFGS [1. 1. 1.] 0.0
BFGS [1. 1. 1.] 0.0
time max =  2
GD [1. 1. 1.] 0.0
GD [1. 1. 1.] 0.0
PGD [1. 1. 1.] 0.0
PGD [1. 1. 1.] 0.0
Nesterov GD [1. 1. 1.] 0.0
Nesterov GD [1. 1. 1.] 0.0
Adagrad GD [1.164 1.169 1.234] 5.312
Adagrad GD [-1.032  1.168  1.232] 6.949
Newton [1. 1. 1.] 0.0
Newton [1. 1. 1.] 0.0
BFGS [1. 1. 1.] 0.0
BFGS [1. 1. 1.] 0.0
Function 3
```

```
n =  2
GD [1.    0.999] 14.188
GD [3.72  2.148] 1557.506
PGD [1.    0.999] 14.18
PGD [3.712 2.105] 1382.062
Nesterov GD [1.    0.999] 14.18
Nesterov GD [3.712 2.105] 1382.072
Adagrad GD [1. 1.] 14.203
Adagrad GD [4.5 4.5] 174808.111
Newton [0. 1.] 14.203
Newton [3.284 3.299] 15043.651
BFGS [1.    0.595] 7.138
BFGS [3.72  2.147] 1553.332
n =  5
GD [1.    0.999] 14.165
GD [3.7   2.032] 1118.108
PGD [1.    0.998] 14.157
PGD [3.694 2.   ] 1016.608
Nesterov GD [1.    0.998] 14.157
Nesterov GD [3.694 2.   ] 1016.614
Adagrad GD [1. 1.] 14.202
Adagrad GD [4.5 4.5] 174803.421
Newton [0. 1.] 14.203
Newton [1.959 2.075] 418.804
BFGS [ 1.001 -0.035] 4.417
BFGS [3.568 1.292] 73.418
n =  10
GD [1.    0.997] 14.127
GD [3.677 1.893] 732.85
PGD [1.    0.997] 14.119
PGD [3.674 1.871] 682.333
Nesterov GD [1.    0.997] 14.119
Nesterov GD [3.674 1.871] 682.336
Adagrad GD [1. 1.] 14.202
Adagrad GD [4.5 4.5] 174797.917
Newton [0. 1.] 14.203
Newton [0.109 1.023] 14.275
BFGS [ 1.002 -0.22 ] 4.357
BFGS [3.456 0.643] 0.126
n =  100
GD [1.    0.973] 13.482
GD [3.605 1.237] 57.122
PGD [1.    0.973] 13.475
PGD [3.605 1.235] 56.311
Nesterov GD [1.    0.973] 13.475
Nesterov GD [3.605 1.235] 56.311
Adagrad GD [1. 1.] 14.198
Adagrad GD [4.5 4.5] 174756.184
Newton [0. 1.] 14.203
Newton [0. 1.] 14.203
BFGS [3.  0.5] 0.0
BFGS [3.  0.5] 0.0
time max =  0.1
GD [1.066 0.665] 7.551
GD [3.617 0.652] 0.087
```

```
PGD [1.253 0.388] 4.03
PGD [3.619 0.621] 0.033
Nesterov GD [1.214 0.43 ] 4.518
Nesterov GD [3.62  0.623] 0.034
Adagrad GD [1. 1.] 14.198
Adagrad GD [4.5 4.5] 174760.64
Newton [0. 1.] 14.203
Newton [0. 1.] 14.203
BFGS [3.  0.5] 0.0
BFGS [3.  0.5] 0.0
time max =  1
GD [1.809 0.11 ] 0.895
GD [3.609 0.619] 0.033
PGD [2.188 0.213] 0.285
PGD [3.587 0.615] 0.031
Nesterov GD [2.174 0.207] 0.3
Nesterov GD [3.597 0.617] 0.032
Adagrad GD [1. 1.] 14.194
Adagrad GD [4.5 4.5] 174711.408
Newton [0. 1.] 14.203
Newton [0. 1.] 14.203
BFGS [3.  0.5] 0.0
BFGS [3.  0.5] 0.0
time max =  2
GD [2.136 0.19 ] 0.343
GD [3.595 0.616] 0.031
PGD [2.497 0.343] 0.074
PGD [3.543 0.608] 0.027
Nesterov GD [2.436 0.32 ] 0.101
Nesterov GD [3.556 0.61 ] 0.028
Adagrad GD [1.001 1.   ] 14.192
Adagrad GD [4.5 4.5] 174692.334
Newton [0. 1.] 14.203
Newton [0. 1.] 14.203
BFGS [3.  0.5] 0.0
BFGS [3.  0.5] 0.0
```

Exercise 6

As the size of element increases it takes more steps to find the minimum. The results show for all equal amount of steps. The best performing method is BFGS. L-BFGS does not give appropriate results, which is probably due to an inaccurate implementation of the method. With a smaller windows size for L-BFGS I got even worse results. As I increased the window size I got more stable results. Below we observe L-BFGS with window size m = 10. Stochastic gives better results then plain GD. Lower RMSE could be achieved if we applied Mini batch Stochastic GD.

```
GD
50 [1.002 0.529] RMSE 0.147
100 [1.233 0.608] RMSE 13.679
1000 [1.985 0.992] RMSE 569.635
10000 [2. 1.] RMSE 5773.504
100000 [2. 1.] RMSE 57735.807
1000000 [2. 1.] RMSE 577351.127
```

```
SGD
50 [0.987 0.961] RMSE 0.244
100 [0.993 0.983] RMSE 0.243
1000 [0.999 0.998] RMSE 0.268
10000 [1. 1.] RMSE 0.269
100000 [1. 1.] RMSE 0.27
1000000 [1. 1.] RMSE 0.269
BFGS
50 [1.    0.486] RMSE 0.077
100 [1.001 0.464] RMSE 0.105
1000 [1.   0.5] RMSE 0.101
10000 [1.   0.5] RMSE 0.1
100000 [1.   0.5] RMSE 0.1
1000000 [1.   0.5] RMSE 0.1
L-BFGS
50 [0.986 0.969] RMSE 0.278
100 [-28.993 825.208] RMSE 1107.076
1000 [ 0.898 69.299] RMSE 34.377
10000 [0.998 0.975] RMSE 10.117
100000 [1.    0.519] RMSE 0.101
1000000 [ 1.   -15.407] RMSE 7.96
```