Machine learning Assignment

MA20278

Harry Lyness

April 28, 2023

Question 4

The value of $t = 32$ gives the fastest convergence. Smaller t is less optimal since slow convergence implies that it may take many iterations to create sufficiently accurate matrices U and W (which is often not very efficient and very time-consuming). Larger t may lead to instability in the method, leading to dramatic increases in loss. The optimal $t = 32$ on average converged in approximately 35 iterations ($\frac{1}{3} * (35 + 31 + 40) \approx 35$).

✓

Proceed to summarise the types of convergence in the loss plots. Note that the loss plots show measure of error of the completed matrix at iteration k and that of the test set. For $t = 1$, observe that the the convergence rate seems to be q-linear but approximately sub-linear,

$$\lim_{k\to\infty} \frac{||L_{\mathcal{I}_{test},Y}(U,W)^{(k+1)}||_2}{||L_{\mathcal{I}_{test},Y}(U,W)^{(k)}||_2} \in (0,1) \approx 1$$

For $t \in \{2, 4, 8, 16, 32\}$, observe that the convergence seems to be $q-$linear,

✓

$$\lim_{k\to\infty} \frac{||L_{\mathcal{I}_{test},Y}(U,W)^{(k+1)}||_2}{||L_{\mathcal{I}_{test},Y}(U,W)^{(k)}||_2} \in (0,1)$$

For $t = 64$, no convergence is observed. This is because there exists a critical iteration k_c such that for increasing $k > k_c$, $L_{\mathcal{I}_{test},Y}(U,W)^{(k)}$ blows up! The average relative error ($||UW^\top - Y||_F / ||Y||_F$) was computed over the three runs.

due to t exceeding loss monotonicity threshold 2/beta

Learning rate	Relative error
1	0.000190
2	0.000188
4	0.000184
8	0.000176
16	0.000159
32	0.000124
64	48.63

If the relative error is low, then $UW^\top$ is accurate at predicting the Y data, and hence the final test loss will also be low. This is because the relative error computed indicates the accuracy of the computed $UW^\top$, and the final test loss is the measure of accuracy between the computed

~error^2

$UW^\top$ and the testing data set. This relationship is clearly observed. For each learning rate $t \in \{1, 2, 4, 8, 16, 32\}$, the final test loss is very low $\approx e{-}9$, corresponding to a very low average relative error (observed for these t values above).

Question 6

$t = 1/16$ gives the fastest convergence. The optimal t for SAG ($t = 32$) is 512 times smaller than the optimal t for SG ($t = 32$). On average, SAG with $t = 1/16$ converged with an average iteration of $\frac{1}{3} * (37125 + 33303 + 31923) = 34177 < K = 50000$, indeed reaching convergence within K iterations.

smaller t is needed due to inexact gradient update

Proceed to summarise the types of convergence in the loss plots. Note that due to stochastic noise, evaluating the convergence rate for $k \to k + 1$ is not that effective. Choose to instead evaluate the convergence rate for $k \to k + n_{iter}$, where n_{iter} is a constant sufficient number of iterations past k. For $t \in \{1, 1/2\}$, observe that for $k < k_{c,t}$ (where $k_{c,t}$ is the critical iteration value at which for increasing $k > k_{c,t}$, $L_{\mathcal{I}_{test},Y}(U, W)$ blows up for each set learning rate $t \in \{1, 1/2\}$), the convergence seems to be q-linear but approximately sub-linear.

$$\lim_{k \to k_{c,t}} \frac{||L_{\mathcal{I}_{test},Y}(U, W)^{(k+n_{iter})}||_2}{||L_{\mathcal{I}_{test},Y}(U, W)^{(k)}||_2} \in (0, 1) \approx 1$$

For $k > k_{c,t}$, the $L_{\mathcal{I}_{test},Y}(U, W)$ blows up. Hence, for $t \in \{1, 1/2\}$, no convergence is observed. For $t \in \{1/4, 1/64\}$, the maximum iteration stopping criterion was met. Therefore, neither $k_{t,c}$ or convergence where observed. The convergence rate observed is approximately sub-linear as $k \to K$. For $t \in \{1/8, 1/16, 1/32\}$ be q-linear/ approximate sub-linear convergence was observed and the sufficient decrease in loss stopping condition was met,

$$\lim_{k \to \infty} \frac{||L_{\mathcal{I}_{test},Y}(U, W)^{(k+n_{iter})}||_2}{||L_{\mathcal{I}_{test},Y}(U, W)^{(k)}||_2} \in (0, 1) \approx 1$$

Note that for all t the convergence rate was predominately just inside $(0, 1)$ only just (occasionally ≥ 1)! This is why convergence is observed for $t \in \{1/8, 1/16, 1/32\}$, but in an extremely large number of iterations. The average relative error ($||UW^\top - Y||_F / ||Y||_F$) of the entire results was computed over the three runs.

Learning rate	Relative error
1	30.0
1/2	4250
1/4	0.000812
1/8	0.000336
1/16	0.000249
1/32	0.000219
1/64	0.00952

Question 7

After re-running the created tests a few times, L_{low} was determined as the $\min_{U,W} L_{\mathcal{I},Y}(U, W)$ across all tests. Therefore, $L_{low} = 0.0017730281702907648 \approx 0.00177$. For GD, $t = 16$ gave the fastest convergence, with an average of $\frac{1}{2}(236 + 595) \approx 416$ iterations (using $\epsilon = 2 * L_{low}$). For SAG, 1/16 gave the fastest convergence, with an average of $\frac{1}{2}(41158 + 30472) = 35815$ iterations

(using $\epsilon = 2 * L_{low}$).

For GD, the rate of convergence for values of $t \in \{1, 2, 4, 8, 16\}$ observed is still approximately sub-linear,

$$\lim_{k\to\infty} \frac{||L_{\mathcal{I}_{test},Y}(U,W)^{(k+1)}||_2}{||L_{\mathcal{I}_{test},Y}(U,W)^{(k)}||_2} \in (1-\tau, 1+\tau) \approx 1 \quad \text{where } \tau << 0.1 \text{ (very small!)}$$

There seem to be osculations in the rate of convergence. This behaviour is certainly observed in the loss plots. Hence, many iterations are required to compute U and W to satisfactory accuracy. However, for increasing t from 1 to 16, the convergence rate does get significantly faster. This is indicated on the loss plots with the increasing steepness of the gradients, allowing for a sufficiently accurate solution to be computed in much fewer iterations (i.e. 3000+ iterations for $t \in \{1, 2\}$ to ≈ 416 for $t = 16$ on average). For $t = 32$, the convergence rate is significantly slower than $t \in \{1, 2\}$, and maximum iteration stopping criterion is met. For $t = 64$, no convergence is observed as $L_{\mathcal{I}_{test},Y}(U,W)^{(k)}$ blows up when $k > k_{c,t}$.

For SAG, observe that there is no convergence for $t \in \{1, 1/2, 1/4, 1/8\}$ since $L_{\mathcal{I}_{test},Y}(U,W)^{(k)}$ blows up when $k > k_{c,t}$. For $t = 1/16$ there is q-linear but approximate sub-linear convergence for rates averaged over $k \to k + n_{iter}$ iterations. For $t \in \{1/32, 1/64\}$ the convergence rate is approximately sub-linear, but the maximum iteration stopping criterion is met. It is interesting how the convergence rate is significantly faster for the first $\approx 0 - 30000$ iterations, and then not much progress is made in reducing the loss for iterations ≥ 30000 (for $t \in \{1/16, 1/32, 1/64\}$, also note similar observations for GD). This observed convergence stagnating occurs at a larger loss for $t \in \{1/32, 1/64\}$ than for $t = 1/16$. Observe that there is more stochastic noise in the convergence rates for larger learning rates as expected.

Tables showing the relative errors for GD and SAG with appropriate inputs.

Learning rate	Relative error
1	0.0688
2	0.0501
4	0.0437
8	0.0437
16	0.0438
32	0.933
64	38.5

Figure 1: A Table to show the average relative error for GD with $t \in \{1, 2, 4, 8, 16, 32, 64\}$, $\epsilon = 2L_{low}$ and $K = 3000$

Learning rate	Relative error
1	54.3
1/2	15.0
1/4	10900
1/8	11.4
1/16	0.0847
1/32	0.112
1/64	0.118

Figure 2: A Table to show the average relative error for SAG with $t \in \{1, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64\}$, $\epsilon = 2L_{low}$ and $K = 50000$

✓ The behaviour of the algorithms is certainly worse for Q7 than for Q4 and Q6. This is predominantly because the tests ran for Q4 and Q6 were computed on a smaller scale and were better conditioned than Q7. For Q4 and Q6, the algorithms converged significantly quicker, on average, with far fewer iterations required to reach the sufficient accuracy stopping criterion (for learning rates with this convergence property...) than for Q7. For Q7 SAG, more learning rates $t \in \{1, 1/2, 1/4, 1/8\}$ experience exploding $L_{\mathcal{I}_{test},Y}(U, W)$ than Q6 ($t \in \{1, 1/2\}$). This certainly indicates worse behaviour. Also, the optimal learning rates were different for GD, with $t = 32$ for Q4 and $t = 16$ for Q7; in fact, the behaviour of Q7 GD with $t = 32$ is worse than $t \in \{1, 2, 4, 8, 16\}$, with large differences observed on the computed loss with increasing iterations k. Interestingly, stagnating at $L_{\mathcal{I}_{test},Y}(U, W)$ for prolonged k iterations was not observed in Q4 and Q6. This is because the theoretical L_{low}, for Q4 and Q6 respectively, were certainly lower than the L_{low} for Q7. This links with the GD results for Q7, it was infeasible to compute results for U and W with a loss of less than $1e-8$. Whereas, this threshold was certainly reached for Q4 GD.

+ since the values y from the exact matrix are not exact values from a rank-r matrix

Other tables for computed average relative error (for Question 7)

Learning rate	Relative error
1	0.0547
2	0.0437
4	0.0449
8	0.0456
16	0.0458
32	0.958
64	38.5

Figure 3: A Table to show the average relative error for GD with $t \in \{1, 2, 4, 8, 16, 32, 64\}$, $\epsilon = 1e-8$ and $K = 5000$

Learning rate	Relative error
1	13.3
1/2	23.0
1/4	29.0
1/8	39000
1/16	0.077
1/32	0.101
1/64	0.118

Figure 4: A Table to show the average relative error for SAG with $t \in \{1, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64\}$, $\epsilon = 1e - 8$ and $K = 50000$