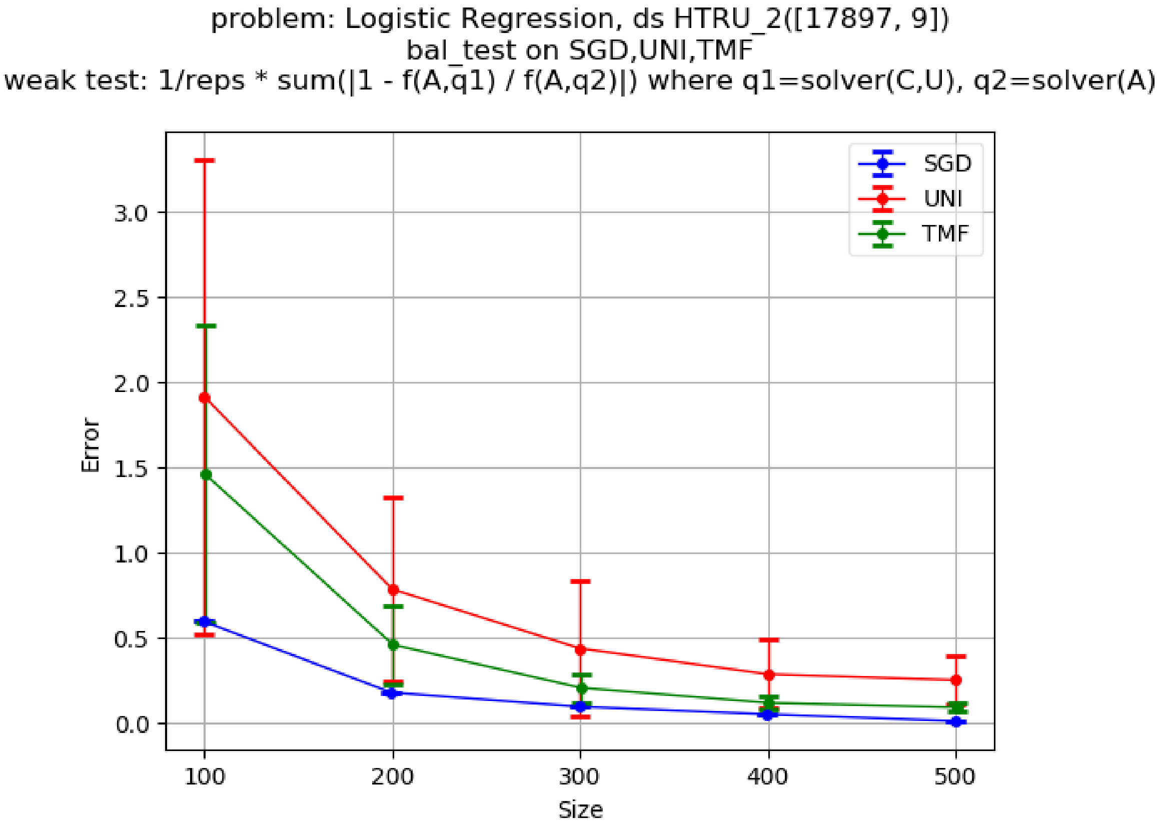
Logistic Regression

- HTRU_2 dataset – size 17897X9
    - Data link
- TMF – coreset made in "Coresets for Near-Convex Functions"
    - link
- Parameters:
    - Repetitions
        - 50 for UNI (Uniform sampling)
        - 1 for SGD (our work)
        - 40 for TMF (external experiment)

weak *test*

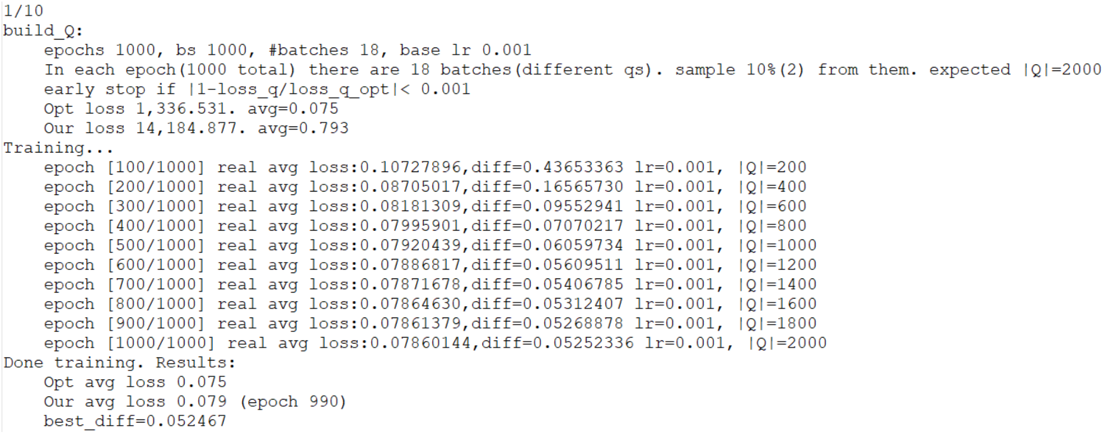
$$y\ values := \frac{1}{repetions} * \sum_{C_i \in Corsets} |1 - f(A, q_1)/f(A, q_2)|$$

$$where\ q_1 = solver(C_i, U_i)\ and\ q2 = solver(A)$$



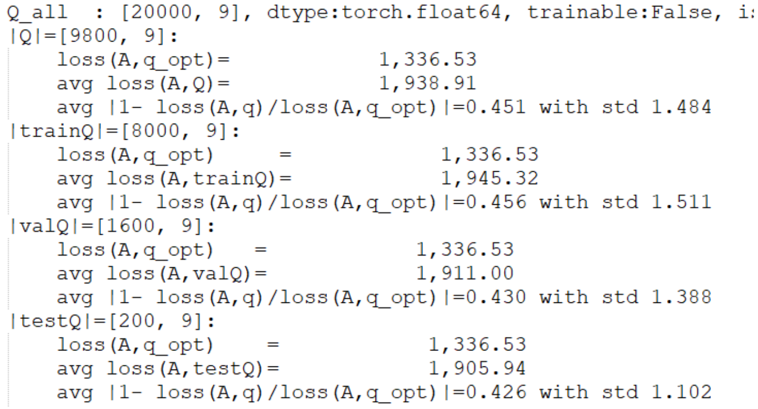
trajectories *test*:

- To create a “real” Q set, we used a simple training process that finds the q opt (the optimal linear regression for this data).
    - We ran the learning process 1000 epochs
    - In each epoch we have #batches states. We sampled 10% of them.
    - We did 10 normal initializations
    - Used Adam optimizer with lr=0.001
    - Example of 1 run out of 10:



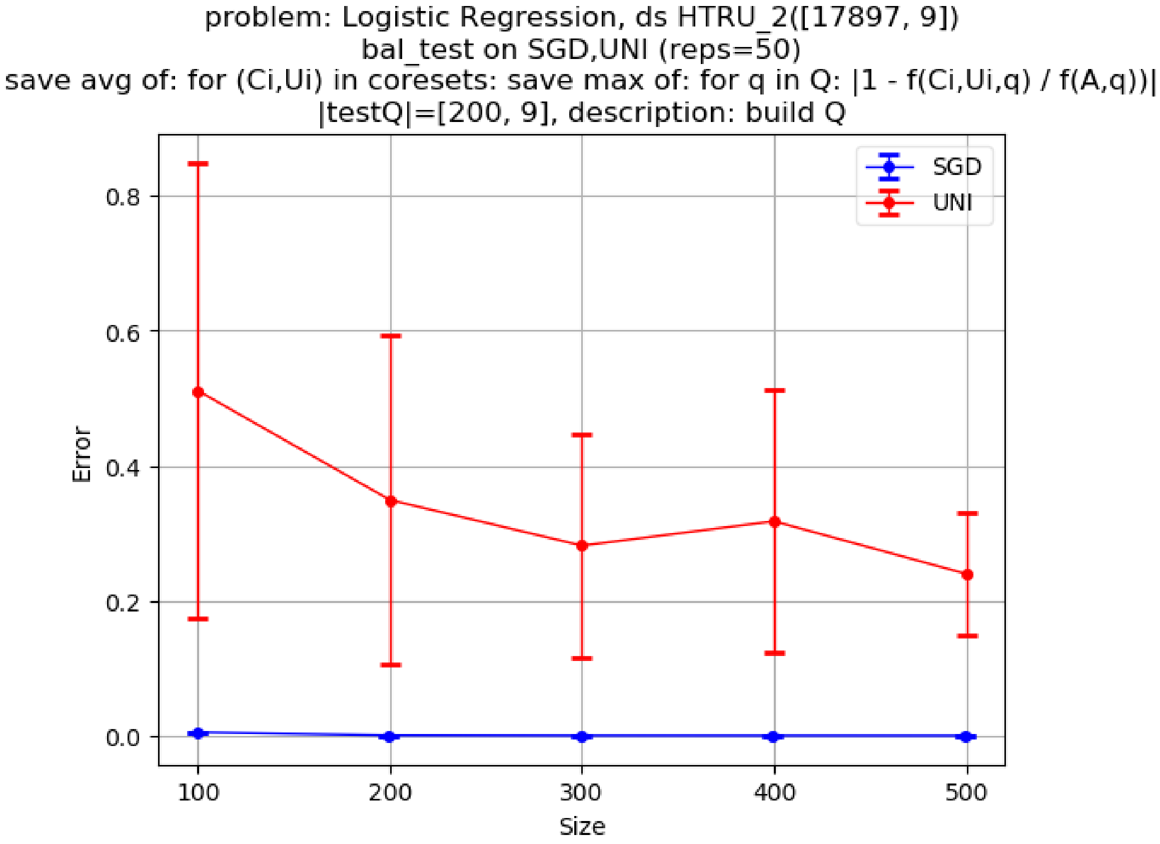
```
1/10
build_Q:
    epochs 1000, bs 1000, #batches 18, base lr 0.001
    In each epoch(1000 total) there are 18 batches(different qs). sample 10%(2) from them. expected |Q|=2000
    early stop if |1-loss_q/loss_q_opt|< 0.001
    Opt loss 1,336.531. avg=0.075
    Our loss 14,184.877. avg=0.793
Training...
    epoch [100/1000] real avg loss:0.10727896,diff=0.43653363 lr=0.001, |Q|=200
    epoch [200/1000] real avg loss:0.08705017,diff=0.16565730 lr=0.001, |Q|=400
    epoch [300/1000] real avg loss:0.08181309,diff=0.09552941 lr=0.001, |Q|=600
    epoch [400/1000] real avg loss:0.07995901,diff=0.07070217 lr=0.001, |Q|=800
    epoch [500/1000] real avg loss:0.07920439,diff=0.06059734 lr=0.001, |Q|=1000
    epoch [600/1000] real avg loss:0.07886817,diff=0.05609511 lr=0.001, |Q|=1200
    epoch [700/1000] real avg loss:0.07871678,diff=0.05406785 lr=0.001, |Q|=1400
    epoch [800/1000] real avg loss:0.07864630,diff=0.05312407 lr=0.001, |Q|=1600
    epoch [900/1000] real avg loss:0.07861379,diff=0.05268878 lr=0.001, |Q|=1800
    epoch [1000/1000] real avg loss:0.07860144,diff=0.05252336 lr=0.001, |Q|=2000
Done training. Results:
    Opt avg loss 0.075
    Our avg loss 0.079 (epoch 990)
    best_diff=0.052467
```

- |Q_all| = 20000
- We sampled |Q| = 9800
    - 8000 for training
    - 1600 for validation
    - 200 for testing



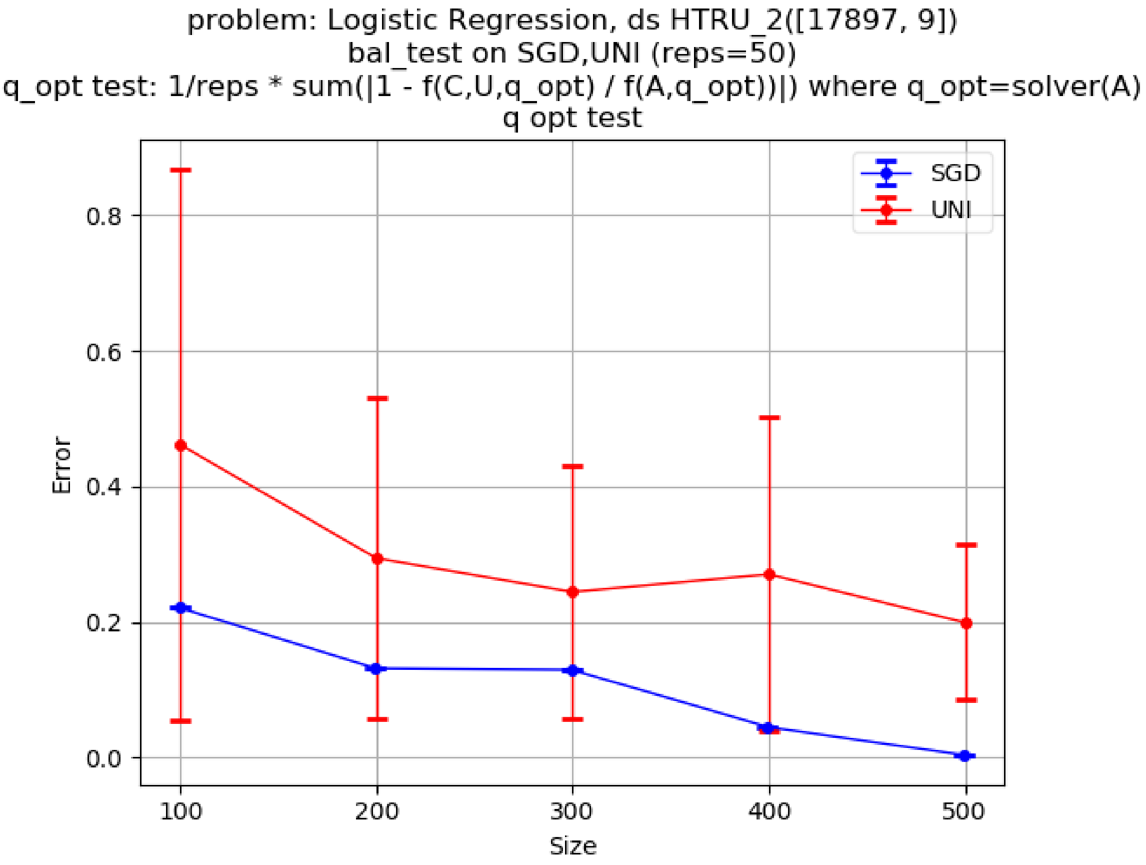
```
Q_all  : [20000, 9], dtype:torch.float64, trainable:False, i:
|Q|=[9800, 9]:
    loss(A,q_opt)=            1,336.53
    avg loss(A,Q)=            1,938.91
    avg |1- loss(A,q)/loss(A,q_opt)|=0.451 with std 1.484
|trainQ|=[8000, 9]:
    loss(A,q_opt)      =            1,336.53
    avg loss(A,trainQ)=             1,945.32
    avg |1- loss(A,q)/loss(A,q_opt)|=0.456 with std 1.511
|valQ|=[1600, 9]:
    loss(A,q_opt)    =            1,336.53
    avg loss(A,valQ)=             1,911.00
    avg |1- loss(A,q)/loss(A,q_opt)|=0.430 with std 1.388
|testQ|=[200, 9]:
    loss(A,q_opt)     =            1,336.53
    avg loss(A,testQ)=             1,905.94
    avg |1- loss(A,q)/loss(A,q_opt)|=0.426 with std 1.102
```

$$y\ values \coloneqq \frac{1}{repetions} * \sum_{C_i \in Corsets} \max_{q_j \in Q} |1 - f(C_i, U_i, q_j)/f(A, q_j)|$$


problem: Logistic Regression, ds HTRU_2([17897, 9])
bal_test on SGD,UNI (reps=50)
save avg of: for (Ci,Ui) in coresets: save max of: for q in Q: |1 - f(Ci,Ui,q) / f(A,q))|
|testQ|=[200, 9], description: build Q
SGD
UNI
Error
0.8
0.6
0.4
0.2
0.0
100
200
300
400
500
Size

q_{opt} $test$

$$y\ values \coloneqq \frac{1}{repetions} * \sum_{C_i \in Corsets} |1 - f(C_i, U_i, q_{opt})/f(A, q_{opt})|\ where\ q_{opt} = solver(A)$$


problem: Logistic Regression, ds HTRU_2([17897, 9])
bal_test on SGD,UNI (reps=50)
q_opt test: 1/reps * sum(|1 - f(C,U,q_opt) / f(A,q_opt))|) where q_opt=solver(A)
q opt test
SGD
UNI
Error
0.0
0.2
0.4
0.6
0.8
100
200
300
400
500
Size