The Role of Adaptive Optimizers for Honest Private Hyperparameter Selection
Shubhankar Mohapatra, Sajin Sasy, Xi He, Gautam Kamath, Om Thakkar
University of Waterloo, Google

Problem
Consider a sensitive dataset D split into a training set and validation set. A trusted curator with a total privacy budget of \((\epsilon, \delta)\) wants to train a model which achieves high accuracy on the validation set. This budget must account for the cost of any queries performed for the sake of hyper parameter selection.

The most popular DP optimizer, DPSGD, has five hyperparameters:
1. Iterations (T)
2. Lot size (L)
3. Learning Rate (\(\alpha\))
4. Clipping Threshold (C)
5. Noise Scale (\(\sigma\))

DPMomentum requires additional momentum tuning \(\epsilon_f, \delta_f\).

Relationship of \(\alpha\) and C
We investigate hyper parameter tuning using Moments Accountant (MA) and Liu and Talwar (LT).

LT is a randomly stopping approach which allows to privately select the best candidate seen by the algorithm from a selected pool of candidates. The random stopping is controlled by the parameter \(\Gamma\). We investigate hyper parameter tuning using Moments Accountant (MA) and Liu and Talwar (LT).

Comparing private optimizers
We compare various private optimizers and show that adaptive optimizer, DPAdam requires the least tuning and achieves consistent performance. Our experiments are performed with constant values of T, L and \(\alpha\).

1. Tuning using DPSGD and DPMomentum requires larger grid search. DPAdam saves tuning of one hyper parameter (\(\alpha\)).
2. The defaults of non-private Adam work for its private counterpart, DPAdam.
3. The best candidates of DPSGD, DPMomentum and DPAdam perform similarly.
4. DPSGD requires higher privacy cost for tuning due to larger candidate pool.
5. DPAdam due to fewer candidates can be tuned using MA for a lesser privacy cost.
6. The learning rate of DPAdam converges after few iterations of training, which we call effective step size (ESS).
7. DPAdamWOSM, a novel optimizer whose learning rate is initialised with ESS starts with this converged learning avoids the second moment computation and enjoys better accuracy at earlier iterations.

Cost of privately tuning DP optimizers

Figure 1: Log of training loss for simulation at \(\sigma = 4\). The white pixels (lowest loss) lie on a diagonal expressing an inverse relationship between LR and C.

Figure 2: Comparing the privacy cost of LT vs MA. The minimal privacy overhead incurred by LT is at least \(\sim 5x\), and increases with the dataset size (left). However, as we allow LT to test more hyper parameters, the privacy cost barely increases (middle). MA is able to test a significant number of candidates at the same cost as the minimal overhead cost of LT (right).

Figure 3: Comparing the testing accuracy curves of different optimisers across their hyper parameter grids with \(\sigma = 4\).

References: