Perturbative Analysis of SGD

0. 2019-03-05

- **Plot** loss curves (predictions vs empirical) for MNIST binary classification ("0" vs "1"). Use super-shallow (logistic non-affine regression) architecture with larger T than before ($T = 100$). We initialize at 0-vector.

- **Tune** explored range of learning rates to be very small ($10^{-6} - 10^{-5}$) to focus on regime wherein empirical generalization gaps scale linearly with learning rate.

- **Observe** discrepancy in plots: theory overestimates generalization gap at 1st order and underestimates benefit of stochasticity at 2nd order. Could this be due to rounding error for very small learning rates?

- **Test** potential that discrepancy was due to rounding error by using 64-bit (instead of 32-bit) floating point precision.

- **Find** no qualitative difference between behaviors with the two precisions.