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.

1. 2019-03-17

- **Modify** loss landscape (motivated by a desire to break symmetry) by initializing at a non-zero constant. Observe a discrepancy at 0th order between predictions and experiment!! This is unexpected indeed!
- **Identify** potential reason for discrepancy: test losses were reported as having incorrectly low variances, leading to the aforementioned 0th order non-overlap.
- **Resolve** discrepancy via two changes: on one hand, estimate test loss variance by comparing losses on different sub-batches of our finite test set; on the other hand, allow for a difference between test-set means and true-distribution means by augmenting relevant error bars by a $\sigma/\sqrt{\text{nb batches in test set}}$ term.
- **Observe** that with these changes, we observe agreement between predicted testscores and experiment (for SGD, GD, and their difference). We also note that for $T = 100$ and $\eta \approx 0.001$ (for this T, this is the natural scale for η given by the critical point of relevant quadratics), the error bars are too big to tell a visual story about the benefit of stochasticity. Alas, Due to the fixity of the aforementioned $\sigma/\sqrt{\text{nb batches in test set}}$ term, running more trials would not clarify that visual story.
- **Fix** mistakes (identified along the way) in uncertainty-visualization, namely: signs for combination of standard deviations (signs should all be positive); and a 0th order term for the uncertainties of generalization gaps that arises from the fact that the actual test-set and train-set means might differ from the true mean.

- **Prioritize** another term to compute in order that, namely the curvature $\mathrm{tr}(H^2)$. This affects the 2nd order behavior of generalization gaps. Will leave for next time...

Also for next time: run on csail machines; start writing notebook to share with dan