EECS 189 Introduction to Machine Learning
Fall 2020

DIS8

This discussion was released **Friday, October 23**.

1 Comparing batch, stochastic, and mini-batch gradient descents

In this discussion, we will observe closely how a neural network learns a function. We will first generate some training data. We will train a network using batch gradient descent, SGD, and mini-batch and will contrast these approaches. We will generate animations that show how well the neural network learns the original data as the network is being trained. Open this notebook in datahub and discuss the questions it contains.

The following problems are 5.b and 5.c from homework 8. They are not dependent on 5.a.

5 SGD in the overparametrized regime

This is a problem that helps you understand why we cared so much about the properties of minimum-norm solutions in the context of machine learning. The standard way of training neural networks in practice is stochastic gradient descent (and variants thereof). We need to understand how it behaves. Here, we just try to minimize squared error. (A similar story holds for logistic loss.)

Consider the standard least-squares problem:

$$\min_{\mathbf{w}} L(\mathbf{w}; \mathbf{X}, \mathbf{y}), \text{ where } L(\mathbf{w}; \mathbf{X}, y) := \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2.$$

Here $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ is a $n \times d$ matrix of features and $\mathbf{y}$ is an n-dimensional vector of labels. Say $d > n$. For a single training sample $(\mathbf{x}_i, y_i)$ let $f_i(\mathbf{w}) := \frac{1}{2}(\mathbf{x}_i^\top \mathbf{w} - y_i)^2$. A *stochastic gradient* is $\nabla_{\mathbf{w}} f_i(\mathbf{w}) = (\frac{d}{d\mathbf{w}} f_i(\mathbf{w}))^\top$ where i is uniformly sampled from $\{1, \ldots, n\}$. By the properties of vector calculus, recall that the derivative of a scalar function of a vector is represented by a row vector, and the gradient is the transpose of that row vector so that we have a regular column vector.

(b) **Prove that any linear combination of stochastic gradients must lie in the row span of X.** That is, it must be a linear combination of the $\{\mathbf{x}_i\}$.

(HINT: Here, you need to actually take the derivative and see what it is.)

(c) Suppose that rows of $\mathbf{X}$ are linearly independent, and that stochastic gradient descent with constant step size $\eta > 0$ initialized at $\mathbf{w}_0 = \mathbf{0}$ converges to something when we view the sequence of vectors $\mathbf{w}_t$ as an infinite sequence. **Show that it converges to the minimum norm interpolating solution $\mathbf{w} = \mathbf{X}^\dagger \mathbf{y}$.** Here $\mathbf{X}^\dagger$ is the Moore-Penrose Pseudoinverse. You can feel free to assume that each training point is used infinitely often by SGD.

(Hint: Remember, what does it mean for a sequence to converge? What does that imply about the individual gradients? What does that mean about the quantity $\|\mathbf{X}\mathbf{w}_t - \mathbf{y}\|^2$*? Finally, this part comes after the previous part for a reason.)*

Contributors:

- Alexander Tsigler
- Anant Sahai
- Inigo Incer