10/27/2021 math notes

Diego Alvarez

October 2021

Stochastic Gradient Descent

We first start with the objective function

$$Q(w) = \frac{1}{n}\sum_{i=1}^{n} Q_i(w)$$

In this case w is trying to minimize $Q(w)$. Each Q_i is an observation of the data.

We are using sum minimization techniques which are similar to certain estimators like the least squares and maximum-likelihood estimation.

Least Squares Estimator

The residuals

$$r_t = y_i - f(x_i, \beta)$$

The Sums

$$S = \sum_{i=1}^{n} r_i^2$$

Maximum Likelihood Estimator

The joint density of samples $X_1, ..., X_n$ that are iid

$$L_n(\theta) = L_n(\theta; y) = f_n(y; \theta)$$

Then maximize the likelihood function over the parametric space

$$\hat{\Theta} = \underset{\theta \in \Theta}{\arg\max}\, \hat{L}_n(\theta; y)$$

Now we want to minimize the function (to each batch if we batch) via the gradient descent method

$$w := w - \eta \Delta Q(w) = w - \frac{\eta}{n}\sum_{i=1}^{n} \nabla Q_i(w)$$