Linear Regression Gradient Descent Derivation

We start with the Linear Regression problem formulation which looks like the following:

$$Y = \Sigma_{i=0}^{d} w_i X_i \tag{1}$$

Maximum Likelihood formulation assuming X and w come from $N(w^T x, \sigma^2)$ distribution:

$$w_{ML} = argmax_w p(y \mid X, w) \tag{2}$$

Considering i.i.d x samples, we expand to:

$$p(y \mid X, w) = \Pi_{i=1}^{n} p(y_i \mid x_i, w) \tag{3}$$

Where $p(y_i \mid x_i, w)$ is just the Normal-Gaussian formula. Now, as explained in the notes, we take the natural log of this function to turn the product into a sum in order to make the formula easier to solve:

$$ln(p(y \mid X, w)) = ln(\Sigma_{i=1}^{n} p(y_i \mid x_i, w)) \tag{4}$$

Further deriving and expanding this, we will observe that a way we can find a w which maximizes or formula (the whole point of this derivation) is by minimizing the following component which is obtained as part of the derivation:

$$Error(w) = (Xw - y)^T (Xw - y) \tag{5}$$

Let's call this $Error(w)$ because this formula resembles the $(predicted - observed)^2$ formulation when plugging in an optimal (converged) w, and this represents squared error.

Gradient Step

The rest is simple, all we must do is find the derivative with respect to w here, also called the gradient and then step in the direction of that gradient to find our optimal w which minimizes this Squared Error.

$$\nabla Error(w) = 2X^T Xw - 2X^T y = X^T (Xw - y) \tag{6}$$

We divide by 2 here because 2 is a constant which we do not care about in our minimization function. Which then makes our gradient step be:

$$w = w - \eta X^T(Xw - y) \tag{7}$$

Where we iterate this process until we have converged to an optimal w. Note here that η (pronounced eta) is a gradient step which controls how quickly we want to step in the direction of optimal w. We must carefully choose this η because if it is too large we may overshoot the optimum and if it is too small it may take too long to converge. Also note that we are descending because our $Error(w)$ function needs to be minimized; however, if there was another function that needed to be maximized, we would instead do $w = w + \eta GradStep$. Furthermore if our function is Convex we are guaranteed to have a global minimum and on the other hand, if it is Concave we are guaranteed to have a global maximum.

Regularization Step

$$Error(w) = (Xw - y)^T(Xw - y) + \lambda w^t w \tag{8}$$

$$\nabla Error(w) = 2X^T Xw - 2X^T y + 2w = X^T(Xw - y) + \lambda w \tag{9}$$

$$w = w - \eta(X^T(Xw - y) + \lambda w) \tag{10}$$

Where we use the L2 Norm here $w^t w$ which ends up penalizing large w-features, thus making our model more robust to outliers and better able to generalize. Finally, λ is used to control the level or efficacy of regularization used (i.e. larger λ means more regularization, and smaller λ means less regularization).