Coordinate Descent

We want to minimize an objective function L. We will work with a single input at a time. Suppose we have $L: \mathbb{R}^d \to \mathbb{R}$ meaning the objective function has d inputs. The algorithm is

1. Initialize $w^{(0)}$
2. Until stopping conditions satisfied
   - Select index i between 1 and d
   - update
     $$w_j^{(t+1)} = w_j^{(t)} \quad \text{for } j \neq i$$
     $$w_i^{(t+1)} \in \underset{z}{\operatorname{argmin}}\, L(w_1^{(t)}, \ldots, z, \ldots, w_d^{(t)})$$

alternatively you could use blocks of indices e.g. pairs of indices

Alternatively we can use guess and check for the update searching over the line in the ith coordinate direction

1. Initialize $w^{(0)}$
2. Until stopping conditions satisfied
   - Select index i between 1 and d
   - Select learning rate $\alpha^{(t)}$

$$w^{(t+1)} = w^{(t)} - \alpha^{(t)} \frac{\partial L}{\partial w_i}(w^{(t)})$$

Suppose L is convex meaning that

$$L(y) = L(x) + DL(x) \cdot (y-x) + \{\text{other terms}\} \quad \overset{0}{\geq}$$

Alternatively

$$L(y) - L(x) \geq DL(x) \cdot (y-x)$$

Here $DL = \left(\frac{\partial L}{\partial x_1}, \ldots, \frac{\partial L}{\partial x_d}\right)$. Think of convex functions as bending upwards from positive second derivatives

Question If L is minimized at $x \in \mathbb{R}^d$ along each coordinate direction, then is x a global minimum of L?

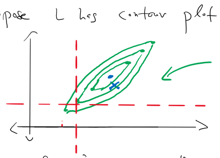
Case 1 Suppose L is differentiable. We have

$$DL = \left(\frac{\partial L}{\partial x_1}(x), \ldots, \frac{\partial L}{\partial x_d}\right) = (0, \ldots, 0)$$

So x is global minimizer by convexity

$$L(y) \geq L(x) + DL(x) \cdot (y-x) = L(x) + 0$$

Case 2 Suppose L has contour plot



the corners are points where the function is not differentiable

Moving away from the corners in the coordinate directions does not reduce the value of the function. Therefore coordinate descent gets stuck at the corners.

Case 3 Can we find a case in between Case 1 and Case 2? Suppose

$$L(x) = g(x) + \sum_{i=1}^{d} h_i(x)$$

where

- g is convex and differentiable
- h_i is convex and possibly not differentiable

Note

$$L(y) - L(x) \geq Dg(x) \cdot (y-x) + \sum_{i=1}^{d} h_i(y) - h_i(x)$$
$$= \sum_{i=1}^{d} \frac{\partial g}{\partial x_i}(x)\,(y_i - x_i) + h_i(y) - h_i(x) \quad \overset{\geq}{0}$$

Therefore coordinate descent works.

Coordinate Descent for Ridge and Lasso

Consider ridge regression with m observations in training set. Take

$$L(w) = |X \cdot w - y|_2^2 + \lambda |w|_2^2$$
$$= \sum_{i=1}^{m} \left(\sum_{j=1}^{d} x_{ij} w_j - y_j\right)^2 + \sum_{j=1}^{d} \lambda w_j^2$$

Note that

$$0 = \frac{\partial L}{\partial w_j} = 2X_{*j}^T (Xw - y) + 2\lambda w_j$$
$$= 2 X_{*j}^T (X_{*j} w_j + X_{*-j} w_{-j} - y) + 2\lambda w_j$$

Therefore

$$w_j = \frac{X_{*j}^T (y - X_{*-j} w_{-j})}{X_{*j}^T X_{*j} + \lambda}$$

Coordinate descent repeats this update for each index.

For lasso regression with m observations in training set

$$L(w) = |X \cdot w - y|_2^2 + \lambda |w|_1$$
$$= \sum_{i=1}^{m} \left(\sum_{j=1}^{d} x_{ij} w_j - y_j\right)^2 + \sum_{j=1}^{d} \lambda |w_j|$$

Note that

$$0 = \frac{\partial L}{\partial w_j} = 2X_{*j}^T (Xw - y) + \lambda \partial |w_j|$$
$$= 2 X_{*j}^T (X_{*j} w_j + X_{*-j} w_{-j} - y) + \lambda \partial |w_j|$$

Therefore

$$w_j = \frac{2 X_{*j}^T (y - X_{*-j} w_{-j}) - \lambda \partial |w_j|}{2 X_{*j}^T X_{*j}}$$
$$\overset{\text{definition}}{=} \frac{c_j - \lambda \partial |w_j|}{a_j}$$

Exercise Show that

$$w_j = \operatorname{sign}\left(\frac{c_j}{a_j}\right) \max\left\{\left|\frac{c_j}{a_j}\right| - \frac{\lambda}{a_j}, 0\right\}$$

This is a threshold function