Shashank Singh[1]

2 LPs and gradient descent in Stats/ML [25 points] (Sashank)

A [4+4+5]

(a) Suppose β optimizes (1). Define

$$\beta_i^+ := \begin{cases} \beta_i : & \beta_i \geq 0 \\ 0 : & \text{else} \end{cases},$$

$\beta^- := \beta^+ - \beta$. Then, $y = X\beta = X(\beta^+ - \beta^-)$ and $\beta^+, \beta^- \geq 0$, so (β^+, β^-) is feasible for (2). Since $1^T(\beta^+ + \beta^-) = \sum_{i=1}^p |\beta_i| = \|\beta\|_1$, the optimum for (2) is at most $\|\beta_1\|$. ■

(b) Suppose (β^+, β^-) optimizes (2). Define $\beta := \beta^+ - \beta^-$. Then, $y = X(\beta^+ - \beta^-) = X\beta$, so β is feasible for (1). Since $\|\beta\|_1 = \sum_{i=1}^p |\beta_i| = 1^T(\beta^+ + \beta^-)$, the optimum for (1) is at most, and therefore equal to, the optimum for (1). ■

(c)

B [6+6]

(a) Rewriting in vector notation (where $h_j(x) = x_j \in \mathbb{R}^n$ is the j^{th} feature vector), we have

$$\hat{\alpha}_j = \operatorname*{argmin}_{\alpha_j \in \mathbb{R}} \|\alpha_j h_j(x) + \hat{y} - y\|_2^2 = \operatorname*{argmin}_{\alpha_j \in \mathbb{R}} \|\alpha_j x_j + \hat{y} - y\|_2^2 = \operatorname*{argmin}_{\alpha_j \in \mathbb{R}} \|\alpha_j x_j - (y - \hat{y})\|_2^2,$$

from which it is apparent that $\hat{\alpha}_j$ is the length of the projection of $y - \hat{y}$ onto x_j,

$$\hat{\alpha}_j = \left\langle \frac{x_j}{\|x_j\|}, y - \hat{y} \right\rangle = \boxed{\langle x_j, y - \hat{y} \rangle.} \tag{1}$$

Note that, rewriting terms as vectors, g is a gradient of the 2-norm recentered at y:

$$g = \frac{\partial L(y, \hat{y})}{\partial \hat{y}} = \frac{\partial \|y - \hat{y}\|_2^2}{\partial \hat{y}} = 2(\hat{y} - y).$$

Thus, rewriting again in vector notation, we have

$$j = \operatorname*{argmin}_{\ell \in \{1,\ldots,M\}} \| - g - \hat{\alpha}_\ell h_\ell(x)\|_2^2 = \operatorname*{argmin}_{\ell \in \{1,\ldots,M\}} \|\hat{\alpha}_\ell x_\ell - (y - \hat{y})\|_2^2.$$

From (1), it is clear that this term is just the error of approximating $(y - \hat{y})$ by its projection onto x_j. This error is minimized by maximizing the inner product of x_j and $y - \hat{y}$, and hence

$$\boxed{j = \operatorname*{argmax}_{\ell \in \{1,\ldots,M\}} |\langle x_j, y - \hat{y} \rangle|.} \tag{2}$$

We could make this derivation a bit more rigorous (find roots of the derivative to compute $\hat{\alpha}_j$, and then obtain (2) via some algebra), but these arguments give much better intuition.

[1] sss1@andrew.cmu.edu

(b)

$$\hat{\alpha}_j = \operatorname*{argmin}_{\alpha_j \in \mathbb{R}} \sum_{i=1}^{n} \log\left(1 + \exp(-2y_i(\hat{y}_i + \alpha_j h_j(x_i)))\right).$$

I don't see a good way of minimizing this analytically. A simple way to approximately minimize this in practice would be to find the α_j values that minimize each term of the sum, and then try values of α_j (perhaps uniformly) in the interval surrounded by those values.