Shashank Singh[1]

4 Regularized logistic regression dual [25 points] (Yifei)

(a) To simplify notation, define $b_i := x_i^T \beta$. For each $i \in \{1, \dots, n\}$, by the choice of y_i,

$$\begin{aligned}\log\left(1 + \exp(-y_i b_i)\right) &= \log\left(\frac{\exp(t_i b_i) + \exp((1 - t_i)b_i)}{\exp(t_i b_i)}\right) \\ &= -\left(t_i b_i - \log\left(\exp(t_i b_i) + \exp((1 - t_i)b_i)\right)\right)\end{aligned}$$

Now observe that, if $t_i = 0$, then $\exp(t_i b_i) = 1$ and $\exp((1 - t_i)b_i) = \exp(b_i)$, while, if $t_i = 1$, then $\exp(t_i b_i) = \exp(b_i)$ and $\exp((1 - t_i)b_i) = 1$. In either case, we have

$$\log\left(1 + \exp(-y_i b_i)\right) = -\left(t_i b_i - \log\left(1 + \exp(b_i)\right)\right). \tag{1}$$

Also, by definition of the 1-norm, clearly

$$\|D\beta\|_1 = \sum_{j=1}^{m} |(D\beta)_j| = \sum_{j=1}^{m} |d_j^T \beta|, \tag{2}$$

and the desired equality follows from (2) and (1). ■

(b) First a quick derivation:

Lemma 1: Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = \log(1 + \exp(x))$, $\forall x \in \mathbb{R}$. Then, the conjugate of f is

$$f^*(y) = y\log(y) + (1 - y)\log(1 - y), \quad \forall y \in [0, 1].$$

Proof: Note that, if

$$0 = \frac{d}{dx} yx - \log(1 + \exp(x)) = y - \frac{\exp(x)}{1 + \exp(x)} = y - \frac{1}{\exp(-x) + 1},$$

and solving for x gives $x = \log\left(\frac{y}{1-y}\right)$. Thus,

$$\begin{aligned}f^*(y) &= \min_{x \in \mathbb{R}} yx - \log(1 + \exp(x)) \\ &= y\log\left(\frac{y}{1 - y}\right) - \log\left(1 + \frac{y}{1 - y}\right) \\ &= y\left(\log(y) - \log(1 - y)\right) + \log(1 - y) = y\log(y) + (1 - y)\log(1 - y). \quad \square\end{aligned}$$

It follows from the relationships between duals, conjugates, and norms (slides 15 and 16 of Lecture 14) that the dual of the given problem is

$$\max_{\alpha \in \mathbb{R}^m} -\sum_{i=1}^{n} (y_i(D^T\alpha)_i)\log(y_i(D^T\alpha)_i) + (1 - y_i(D^T\alpha)_i)\log(1 - y_i(D^T\alpha)_i),$$

such that $\|\alpha\|_\infty \leq \lambda, 0 \leq y_i(D^T\alpha)_i \leq 1$ (since $(D\beta)^T\alpha = \beta^T(D^T\alpha)$). ■

[1] sss1@andrew.cmu.edu

(c) Since the regularization term of the original problem is not generally smooth, we could only use the subgradient method or generalized gradient descent. The dual, on the other hand, is smooth, but has inequality constraints, so we could use gradient descent. In neither case could we use Newton's or quasi-Newton methods, or conjugate gradient descent.

(d) As usual, the solution to the dual problem lower bounds the solution to the primal problem. Didn't have time to finish this.