2.10 Bibliographic notes

Appendix

Convex optimization with constraints

In this Appendix we present an overview of convex optimization concepts that are particularly useful for the lasso and other methods described later in this book.

Suppose that $f_0(\beta)$ and $f(\beta)$ are convex functions of a vector argument $\beta \in R^p$ and we want to solve the following constrained optimization problem:

$$\text{minimize} f_0(\beta) \;\; \text{subject to} \;\; f(\beta) = 0 \tag{2.19}$$

The method of *Lagrange multipliers* solves this problem by solving the system of equations

$$\bigtriangledown f_0(\beta) + \lambda \bigtriangledown f(\beta) = 0 \tag{2.20}$$

over β and λ. In other words, the solution (β^*, λ^*) satisfies $\bigtriangledown f_0(\beta^*) = -\lambda^* \bigtriangledown f(\beta^*)$. Geometrically, this says that at β^*, the tangent vectors of the contour lines of f_0 and f are parallel, or equivalently the norm vector of $f_0(\beta)$ is at right angles to the tangent of the constraint contour $f(\beta) = 0$. This means that if we are at the point β^* and travel along the contour $f(\beta) = 0$, we cannot decrease the value of $f_0(\beta)$. See Figure 2.7 for an example.

Now we consider the more general problem:

$$\text{minimize} f_0(\beta) \;\; \text{subject to} \;\; f_k(\beta) \leq 0, k = 1, 2, \ldots m \tag{2.21}$$

Since $f_0(\beta)$ and $f_k(\beta)$ are convex, there is a unique global minimum. Suppose first that each $f_k(\beta)$ is differentiable. Then necessary conditions for β^* to be the global solution are as follows: there exists a $\lambda^* \geq 0$ such that

1. $f(\beta) \leq 0$ (that is, β is a feasible point for the constraints)
2. $\lambda_k^* f_k(\beta) = 0, k = 1, 2, \ldots m$ (complementary slackness)
3. $\bigtriangledown f_0(\beta^*) + \sum_{k=1}^m \lambda_k^* \bigtriangledown f_k(\beta *) = 0$

Here λ_j^* are the Lagrange multipliers or *dual* variables in the problem, while β is called the *primal* variable. These conditions are known as the *Karush-Kuhn-Tucker* optimality conditions. They are also sufficient conditions for a global minimum under regularity assumptions on f and $\{f_k\}$ known as

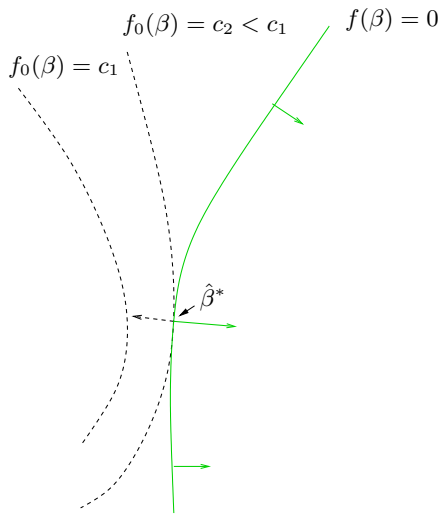

FIGURE 2.7.

"strong duality". These conditions generalize the usual Lagrange multiplier method discussed above. With inequality constraints, condition (2) says that the Lagrange multiplier λ_k^* must be zero if the constraint $f_k(\beta) \leq 0$ is inactive, that is, if $f_k(\beta)$ is strictly < 0. This says if the constraint $f_k(\beta) \leq 0$ is not enforced (tight) at the solution β^*, then the contour of $f_k(\beta)$ is not tangent to the contour of $f(\beta)$ at β^*. The gradients must also point in opposite directions ($\lambda_k^* > 0$) for the point β^* to be a minimum.

There exists many numerical methods for solving this problem, for example interior point logarithmic barrier techniques. In most of the problems in this book, however, the constraint $f(\beta)$ is not differentiable and so we can not directly apply the KKT conditions to characterize the solution. In the lasso, for example, $f(\beta) = \sum_j |\beta_j| - t$ and this function is not differentiable if any of the β_j are equal to 0. One way to finesse this is to write each β_j in terms of its positive and negative parts:

$$\beta_j = \theta_j^+ - \theta_j^-;\ j = 1, 2, \ldots p \tag{2.22}$$

with $\theta_j^+, \theta_j^- \geq 0$. Then $|\beta_j| = \theta_j^+ + \theta_j^-$ and we reparametrize the problem in terms of the θ_j^+, θ_j^-, $j = 1, 2, \ldots p$. The new problem has differentiable constraints and so the KKT conditions apply. However this approach doubles the number of parameters and increases the number of constraints from 1 to $2p + 1$.

A more streamlined approach uses the notion of *subgradients*. A vector g is called a subgradient of a convex function f at β if

$$f(b) \geq f(\beta) + g^T(b - \beta) \tag{2.23}$$

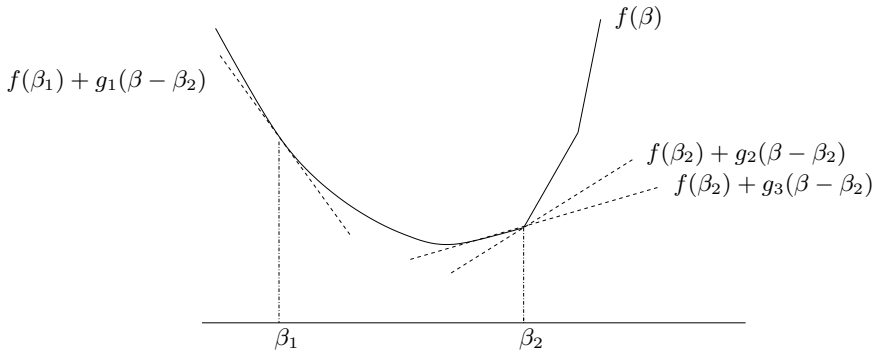

FIGURE 2.8.

for all b in the domain of f. This says that $g^T(b - \beta)$ is a non-vertical supporting hyperplane to $f(\beta)$ at β. The vector g is the normal vector to this hyperplane.

Figure 2.8 shows a function $f(\beta)$ and some subgradients at two points β_1 and β_2. At β_1 the function is differentiable and hence there is only one subgradient. At β_2 it is not differentiable and the subgradients are all lines that touch the function at β_2 and lie below it.

The set of all subgradients of f at β is called the *subdifferential* and is denoted by $\partial f(\beta)$. For example if $f(\beta) = |\beta|$, then $\partial f(\beta) = 1$ if $\beta > 1$, $\partial f(\beta) = -1$ if $\beta < 1$ and $\partial f(\beta) = [-1, 1]$ if $\beta = 0$.

How is this useful? In the case where the functions f_0 or f_j are non-differentiable, the theory of subgradients says that the KKT conditions apply, but with condition (3) above replaced by

$$0 \in \partial f_0(\beta^*) + \sum_{k=1}^{m} \lambda_k^* \partial f_k(\beta *) \tag{2.24}$$

That is, we simply replace the gradients in KKT condition (3) by subdifferentials.

Using this, we can solve our minimization problem by find a solution to these subgradient equations (2.24). For example, if $f(\beta) = \sum_j |\beta_j| - t$, and f_0 is differentiable, then $\partial f(\beta) = \sum_j \partial f(\beta_j)$ and (2.24) becomes

$$\bigtriangledown f_0(\beta^*) + \lambda^* s(\beta) = 0 \tag{2.25}$$

where $s(\beta) = (s_1, s_2, \ldots s_p)$, $s_j = \text{sign}(\beta)$ if $\beta \neq 0$ and $s_j \in [-1, 1]$ if $\beta = 0$. This is sometimes written as $s_j \in \text{sign}(\beta)$. When $f_0(\beta)$ is squared loss, this gives Equation (2.11).

Separability and coordinate descent

Consider minimization of a function of the form

$$J(\beta_1, \ldots \beta_p) = g(\beta_1, \ldots \beta_p) + \sum_{i=j}^{p} h_j(\beta_j) \tag{2.26}$$

where $g(\cdot)$ is differentiable and convex, and the $h_j(\cdot)$ are convex. Here each β_j can be a vector, but the different vectors cannot have any overlapping members.

A *coordinate descent algorithm* successively minimizes the function over each parameter, holding the others fixed:

$$\hat{\beta}_j \leftarrow \text{argmin}_{\beta_j} J(\hat{\beta}_1, \hat{\beta}_2 \cdots \hat{\beta}_{j-1}, \beta_j, \hat{\beta}_{j+1} \cdots \hat{\beta}_p) \tag{2.27}$$

for $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$. After each minimization, we replace β_j by $\hat{\beta}_j$ and then move on to the next coordinate.

Tseng (1988) (see also Tseng (2001)) shows that coordinate descent converges to the minimizer of J. The key to this result is the separability of the penalty function $\sum_{j=1}^{p} h_j(\beta_j)$, a sum of functions of each individual parameter. This result implies that the coordinate-wise algorithms for the lasso, (and other problem discussed in this book) converge to their optimal solutions.

When the penalty is not separable, coordinate descent can get "stuck" and not reach the global minimum. An example is the fused lasso (Chapter xxx) where the penalty has the form $\sum_{j=1}^{n} |\beta_j - \beta_{j-1}|$.

References

Tseng, P. (1988), Coordinate ascent for maximizing nondifferentiable concave functions, Technical Report LIDS-P ; 1840, Massachusetts Institute of Technology. Laboratory for Information and Decision Systems.

Tseng, P. (2001), 'Convergence of block coordinate descent method for nondifferentiable maximization', *J. Opt. Theory and Applications* **109**(3), 474–494.