APM462H1S: Nonlinear optimization, Winter 2014.

Summary of February 3 lecture.

The February 3 lecture covered a selection of material from sections 9.1 - 9.3 of the textbook.

In particular, we went over the complete proof of the *conjugate directions* theorem from Section 9.1, and we also discussed (without complete proofs) the *expanding subspace* theorem in Section 9.1, and the basics of the *conjugate gradient* method from section 9.3.

Some main points are:

- The conjugate directions method gives an efficient procedure for minimizing quadratic functions
$$f(x) = \frac{1}{2}x^T Q x + b^T x$$
for $x \in E^n$ where Q is a symmetric, positive definite $n \times n$ matrix. In order to carry it out, however, one needs to have a set of n nonzero Q-conjugate vectors $d_0, \ldots, d_{n-1}$. That is, these vectors are supposed to satisfy
$$d_i^T\, Q\, d_j = 0 \qquad \text{whenever } i \neq j.$$
- The conjugate gradient method is similar, but it provides a way of generating the vectors $d_0, \ldots, d_{n-1}$ as the algorithm runs, so that it is not necessary to find the set of Q-orthogonal directions before starting the minimization procedure.
- you do not need to remember the exact formulas for either method, eg, for conjugate directions,
$$x_{k+1} = x_k + \alpha_k d_k, \qquad \alpha_k = -\frac{d_k^T g_k}{d_k^T Q d_k}, \qquad g_k = Qx_k - b$$
but you should remember the idea: x_{k+1} has the form $x_k + \alpha_k d_k$, where α_k is the minimizer of $\phi(\alpha) = f(x_k + \alpha d_k)$. Ideally, you should be able to start from the general idea and figure out the exact formula for α_k rather quickly.)
- Similarly, for the conjugate gradient method, you should remember
$$x_{k+1} = \text{ exactly as for conjugate directions}$$
but d_k is found by starting with g_k and "correcting it" to make it be Q-orthogonal to d_{k-1}:
$$d_k = g_k + \beta_{k-1} d_{k-1}, \qquad \beta_{k-1} \text{ chosen so that } d_{k-1}^T Q d_k = 0.$$
(For $k = 1$, we just take $d_0 = g_0$.)
- Although you do not need to memorize the formula for α_k, you certainly *are welcome* to remember it if you feel like it. (It's a pretty useful formula, and in particular, a very similar formula also arises in the method of steepest descent, where however $d_k = g_k$.)
- Both the conjugate gradient and conjugate directions methods are *guaranteed* to converge to the actual minimizer in at most n steps, for quadratic minimization problems in E^n. This is much better than the method of steepest descent.

- make sure that you are comfortable with the whole idea of vectors that are conjugate orthogonal. Later on I will suggest some practice exercises about this point, if you want further practice.
- in particular, please make sure that you really understand why a set of nonzero Q-orthogonal vectors is linearly independent (this was proved in the lecture, and the proof can be found in Section 9.1 of the book). Also make sure that you understand the very similar argument that shows why, if
$$x^* - x_0 = \alpha_0 d_0 + \cdots + \alpha_{n-1} d_{n-1}$$
then the coefficients α_i must have the form
$$\alpha_i = \frac{d_i^T Q(x^* - x_0)}{d_i^T Q d_i}.$$

Example: Here is a concrete example of the conjugate gradient method, which was presented, with some details missing, at the beginning of the lecture.

Consider the function f, defined on E^2 by
$$f(x,y) = \frac{1}{2}x^2 + 50y^2.$$
This can also be written
$$f(x,y) = \frac{1}{2}\vec{x}^T Q \vec{x}, \qquad \text{for } Q = \begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix}, \qquad \vec{x} = \begin{bmatrix} x \\ y \end{bmatrix}.$$

On January 27, we saw that if we if we start at an unfavorable initial guess (x_0, y_0) and try to minimize this by the method of steepest descent, then it converges rather poorly. In particular, if we start at $(x_0, y_0) = (100, 1)$ then we saw that
$$(x_1, y_1) = \frac{99}{101}(100, -1)$$
$$(x_2, y_2) = (\frac{99}{101})^2(100, 1)$$
$$(x_3, y_3) = (\frac{99}{101})^3(100, -1)$$

and generally,
$$(x_k, y_y) = (\frac{99}{101})^k(100, (-1)^k).$$
Thus, for the particular initial guess that we considered, the method converges pretty slowly.

Now let's consider the conjugate gradient method, starting at the same point $\vec{x}_0 = (x_0, y_0) = (100, 1)$.

Then $d_0 = g_0 = Q\vec{x}_0 = (100, 100)$, so
$$\alpha_0 = -\frac{d_0^T g_0}{d_0^T Q d_0} \;=\; -\frac{2}{101}$$
and
$$\vec{x}_1 = \vec{x}_0 + \alpha_0 d_0 = (100, 1) - \frac{200}{101}(1, 1) = \frac{99}{101}(100, -1)$$

as with the method of steepest descent (which starts out identical to the conjugate gradient method, on the first step). Next,

$$g_1 = Q\,\vec{x}_1 = \frac{99}{101}(100, -100)$$

so using the general formula, we get

$$d_1 = g_1 + \beta_0 d_0, \qquad \beta_0 = -\frac{d_0^T Q g_1}{d_0^T Q d_0} = (\frac{99}{101})^2.$$

Thus some calculations lead to:

$$d_1 = \frac{99}{101}(100, -100) + (\frac{99}{101})^2(100, 100) = \frac{9900}{101}\frac{2}{101}(100, -1).$$

It is a bit unpleasant to do all these computations by hand. But something remarkable[1] happens: the vector d_1 points exactly parallel to the line joining $\vec{x}_1$ and the origin, which is the global minimum of f. So when we minimize f along the line passing through $\vec{x}_1$ and parallel to d_1, the minimum will occur at the origin. So

$$\vec{x}_2 = (0, 0) = \text{ the global minimum.}$$

We could also figure this out by blindly applying the formula $\vec{x}_2 = \vec{x}_1 + \alpha_1 d_1$, for the right value of α_1, but there is no need, since it is clear what the answer will be.

Thus, as predicted by our general theory, the conjugate gradient method converges after 2 steps (for this problem, where we are minimizing a quadratic function of 2 variables). Although we had to work a little harder at each step, altogether this is a great improvement over the method of steepest descent.

[1] that is, it looks remarkable to us if we forget the underlying reason why it *has* to happen, which we already know about.