PDCO: PRIMAL-DUAL METHOD FOR OPTIMIZATION WITH CONVEX OBJECTIVES

MICHAEL SAUNDERS, October 23, 2002

NOTES FOR DERIVING THE EQUATIONS

1. The Problem. The nominal optimization problem is

$$\boxed{\begin{array}{lll} \text{NP} \qquad\qquad & \underset{x}{\text{minimize}} & \phi(x) \\ & \text{subject to} & Ax = b, \quad \ell \le x \le u, \end{array}}$$

where $\phi(x)$ is a convex function with known gradient $g(x)$ and Hessian $H(x)$, and $A \in \mathbb{R}^{m\times n}$. The format of NP is suitable for any linear constraints. For example, a double-sided constraint $\alpha \le a^T\widetilde{x} \le \beta$ $(\alpha < \beta)$ should be entered as $a^T\widetilde{x} - \xi = 0$, $\alpha \le \xi \le \beta$, where $\widetilde{x}$ and ξ are relevant parts of x.

To allow for constrained least-squares problems, and to ensure unique primal and dual solutions (and improve solver stability), we regularize the problem as

$$\boxed{\begin{array}{lll} \text{NP2} \qquad\qquad & \underset{x,r}{\text{minimize}} & \phi(x) + \frac{1}{2}\|D_1x\|^2 + \frac{1}{2}\|r\|^2 \\ & \text{subject to} & Ax + D_2r = b, \quad \ell \le x \le u, \end{array}}$$

where D_1, D_2 are positive-definite diagonal matrices specified by the user. The diagonals of D_1 are typically small (10^{-3} or 10^{-4}). Similarly for D_2 if the constraints in NP should be satisfied reasonably accurately. For least-squares applications, some of the diagonals of D_2 will be 1. Note that some elements of ℓ and u may be $-\infty$ and $+\infty$ respectively, but we expect no large numbers in A, b, D_1, D_2. If $\|D_2\|$ is small, we would expect A to be under-determined $(m < n)$. If $D_2 = I$, A may have any shape.

2. The Barrier Approach. First we introduce slack variables x_1, x_2 to convert the bounds to non-negativity constraints:

$$\boxed{\begin{array}{lll} \text{NP3} \qquad\qquad & \underset{x,r,x_1,x_2}{\text{minimize}} & \phi(x) + \frac{1}{2}\|D_1x\|^2 + \frac{1}{2}\|r\|^2 \\ & \text{subject to} & \begin{array}{rl} Ax + D_2r &= b \\ x - x_1 &= \ell \\ x + x_2 &= u \\ x_1, x_2 &\ge 0. \end{array} \end{array}}$$

Then we replace the non-negativity constraints by the log barrier function, obtaining a sequence of convex subproblems with decreasing values of μ $(\mu > 0)$:

$$\boxed{\begin{array}{lllll} \text{NP}(\mu) & \underset{x,r,x_1,x_2}{\text{minimize}} & \multicolumn{3}{l}{\phi(x) + \frac{1}{2}\|D_1x\|^2 + \frac{1}{2}\|r\|^2 - \mu\sum_j \ln([x_1]_j[x_2]_j)} \\ & & Ax + D_2r = & b & : y \\ & \text{subject to} & x - x_1 = & \ell & : z_1 \\ & & -x - x_2 = & -u, & : z_2 \end{array}}$$

where y, z_1, z_2 denote dual variables for the associated constraints. With $\mu > 0$, most variables are strictly positive: $x_1, x_2, z_1, z_2 > 0$.

The KKT conditions for the barrier subproblem involve the three *primal* equations of NP(μ), along with four *dual* equations stating that the gradient of the subproblem objective should be a linear combination of the gradients of the primal constraints:

$$\begin{aligned} Ax + D_2 r &= \;\; b \\ x - x_1 &= \;\; \ell \\ -x - x_2 &= -u \\ A^T y + z_1 - z_2 &= \;\; g(x) + D_1^2 x & &: x \\ D_2 y &= \;\; r & &: r \\ X_1 z_1 &= \;\; \mu e & &: x_1 \\ X_2 z_2 &= \;\; \mu e, & &: x_2 \end{aligned}$$

where $X_1 = \operatorname{diag}(x_1)$, $X_2 = \operatorname{diag}(x_2)$, and similarly for Z_1, Z_2 later. Actually the last two equations arise in a different form. The dual equation for x_1 is really

$$-z_1 = \nabla(-\mu \ln(x_1)) = -\mu X_1^{-1} e,$$

where e is a vectors of 1's. Thus, $x_1 > 0$ implies $z_1 > 0$, and multiplying by $-X_1$ gives the equivalent equation $X_1 z_1 = \mu e$ as stated. In this form, the last two equations are commonly called (perturbed) *complementarity* conditions.

3. Newton's Method. We now eliminate $r = D_2 y$ and apply Newton's method:

$$\begin{aligned} A(x + \Delta x) + D_2^2 (y + \Delta y) &= \;\; b \\ (x + \Delta x) - (x_1 + \Delta x_1) &= \;\; \ell \\ -(x + \Delta x) - (x_2 + \Delta x_2) &= -u \\ A^T (y + \Delta y) + (z_1 + \Delta z_1) - (z_2 + \Delta z_2) &= \;\; g + H\Delta x + D_1^2 (x + \Delta x) \\ X_1 z_1 + X_1 \Delta z_1 + Z_1 \Delta x_1 &= \;\; \mu e \\ X_2 z_2 + X_2 \Delta z_2 + Z_2 \Delta x_2 &= \;\; \mu e, \end{aligned}$$

where g and H are the current objective gradient and Hessian. To solve this Newton system, we work with three sets of residuals:

$$\begin{pmatrix} \Delta x - \Delta x_1 \\ -\Delta x - \Delta x_2 \end{pmatrix} = \begin{pmatrix} r_\ell \\ r_u \end{pmatrix} \equiv \begin{pmatrix} \ell - x + x_1 \\ -u + x + x_2 \end{pmatrix}, \tag{1}$$

$$\begin{pmatrix} X_1 \Delta z_1 + Z_1 \Delta x_1 \\ X_2 \Delta z_2 + Z_2 \Delta x_2 \end{pmatrix} = \begin{pmatrix} c_\ell \\ c_u \end{pmatrix} \equiv \begin{pmatrix} \mu e - X_1 z_1 \\ \mu e - X_2 z_2 \end{pmatrix}, \tag{2}$$

$$\begin{pmatrix} A\Delta x + D_2^2 \Delta y \\ -H_1 \Delta x + A^T \Delta y + \Delta z_1 - \Delta z_2 \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \equiv \begin{pmatrix} b - Ax - D_2^2 y \\ g + D_1^2 x - A^T y - z_1 + z_2 \end{pmatrix}, \tag{3}$$

where $H_1 = H + D_1^2$. We use (1) and (2) to replace two sets of vectors in (3). With

$$\begin{pmatrix} \Delta x_1 \\ \Delta x_2 \end{pmatrix} = \begin{pmatrix} -r_\ell + \Delta x \\ -r_u - \Delta x \end{pmatrix}, \qquad \begin{pmatrix} \Delta z_1 \\ \Delta z_2 \end{pmatrix} = \begin{pmatrix} X_1^{-1}(c_\ell - Z_1 \Delta x_1) \\ X_2^{-1}(c_u - Z_2 \Delta x_2) \end{pmatrix}, \tag{4}$$

$$\begin{aligned} H_2 &\equiv H + D_1^2 + X_1^{-1} Z_1 + X_2^{-1} Z_2 \\ w &\equiv r_2 - X_1^{-1}(c_\ell + Z_1 r_\ell) + X_2^{-1}(c_u + Z_2 r_u) \end{aligned} \tag{5}$$

we find that

$$\begin{pmatrix} -H_2 & A^T \\ A & D_2^2 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} w \\ r_1 \end{pmatrix}. \tag{6}$$

4. Solving for $(\Delta x, \Delta y)$. If $\phi(x)$ is a general convex function with known Hessian H, system (6) may need to be treated directly by sparse or iterative methods.

Alternatively it may be possible to obtain a sparse Cholesky factorization

$$H_2 = LL^T \quad \text{or} \quad H_2 \approx LL^T,$$

where $H_2 = H +$ diagonal terms, and L is a nonsingular permuted triangle. This is trivial if $\phi(x)$ is a separable function, since H and H_2 in (5) are then diagonal. In other cases it may suffice to use $\mathrm{diag}(H)$ or some other approximation to H in the definition of H_2 and L, thereby implementing a pseudo-Newton method for obtaining reasonable directions $(\Delta x, \Delta y)$.

System (6) may now be solved by eliminating either Δx or Δy:

$$(A^T D_2^{-2} A + H_2)\Delta x = A^T D_2^{-2} r_1 - w, \qquad D_2^2 \Delta y = r_1 - A\Delta x, \tag{7}$$

or

$$(A H_2^{-1} A^T + D_2^2)\Delta y = A H_2^{-1} w + r_1, \qquad H_2 \Delta x = A^T \Delta y - w. \tag{8}$$

Sparse Cholesky factorization may again be applicable to the left-hand parts of these systems, but for numerical reasons it is preferable to regard them as least squares problems:

$$\min_{\Delta x} \left\| \begin{pmatrix} D_2^{-1} A \\ L^T \end{pmatrix} \Delta x - \begin{pmatrix} D_2^{-1} r_1 \\ -L^{-T} w \end{pmatrix} \right\|^2, \qquad D_2 \Delta y = D_2^{-1}(r_1 - A\Delta x), \tag{9}$$

or

$$\min_{\Delta y} \left\| \begin{pmatrix} L^{-1} A^T \\ D_2 \end{pmatrix} \Delta y - \begin{pmatrix} L^{-1} w \\ D_2^{-1} r_1 \end{pmatrix} \right\|^2, \qquad L^T \Delta x = L^{-1}(A^T \Delta y - w). \tag{10}$$

The right-most vectors in (9)–(10) are part of the residual vectors for the least squares problems (and may be by-products from the least squares computation).