Enriched recursive multi-level optimization

N. I. M. Gould, M. Kočvara, D. P. Robinson, & Ph. L. Toint

Working note RAL-NA-2009–2

1st October 2009

1 Iterated subspace minimization

We consider the problem of minimizing the twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. Given an estimate x_k of the required solution we select

$$x_{k+1} = \arg\min_{y \in \mathbb{R}^{n_k}} f(x_k + S_k y), \tag{1.1}$$

where $1 \leq n_k \leq n$ and S_k is a selected n by n_k subspace matrix. The convergence of methods based upon this simple idea is considered by [1, 4, 11, 12, 13, 14]. In particular, the minimization in (1.1) needs only be performed approximately for the iterates to converge to a critical point of f.

One possibility is to construct the subspace matrix $S_k = (s_{k,1} \cdots s_{k,n_k})$ as follows. Consider the second-order model $m_k(s) \stackrel{\text{def}}{=} s^T g_k + \frac{1}{2} s^T H_k s$, where $g_k \stackrel{\text{def}}{=} g(x_k)$ and $H_k \stackrel{\text{def}}{=} H(x_k)$, and where $g(x)$ and $H(x)$ are respectively the gradient and Hessian of f at x. Let $\Delta_{k,i} > 0$ for $1 \leq i \leq n_k$, and let $\mathcal{T}(\Delta) \stackrel{\text{def}}{=} \{s \mid \|s\|_2 \leq \Delta\}$ for given $\Delta > 0$. Then select

$$s_{k,1} = \arg\min_{s \in \mathcal{T}(\Delta_{k,1})} m_k(s), \tag{1.2}$$

and more generally

$$s_{k,i} = \arg\min_{\substack{s^T s_{k,j} = 0,\, j<i, \\ s \in \mathcal{T}(\Delta_{k,i})}} m_k(s) \tag{1.3}$$

for $2 \leq i \leq n_k$. The problem (1.2) is the well-know trust-region subproblem, and there are many efficient methods for its solution [3, 5, 6, 7, 8, 9, 10]. The generalisation (1.3) may also be solved using these methods by implicitly constraining the solution into the subspace $\{s \mid s^T s_{k,j} = 0, j < i\}$ by (for instance) projection [3, §5].

Many variants are possible. For instance the gradient direction g_k may be added to the subspace, in which case the solution to (1.1) will satisfy $g_k^T g_{k+1} = 0$.

2 Convergence of iterated subspace minimization

Suppose that we have built an n by n_k subspace matrix S_k, at least one of whose columns—say the first, $s_{k,1}$—has a non-trivial component in the gradient direction, in the sense that

$$|s_{k,1}^T g_k| \geq \kappa_{\rm g} \|g_k\| \tag{2.4}$$

for some $\kappa_{\rm g} \in (0,1]$. Let

$$f_k^{\rm S}(y) \stackrel{\rm def}{=} f(x_k + S_k y)$$

be the objective function centered at x_k and restricted to the affine space $x_k + S_k y$, and let

$$g_k^{\rm S}(y) \stackrel{\rm def}{=} \nabla_y f_k^{\rm S}(y) = S_k^T g(x_k + S_k y) \;\text{ and }\; H_k^{\rm S}(y) \stackrel{\rm def}{=} \nabla_{yy} f_k^{\rm S}(y) = S_k^T H(x_k + S_k y) S_k$$

be its gradient and Hessian respectively.

Now suppose that we generate a new iterate $x_{k+1} = x_k + S_k y_k$, so that y_k improves $f_k^{\rm S}(y)$ in as much as

$$f(x_{k+1}) \leq f(x_k) - \gamma_{\rm g} \|g_k\| \min[\|g_k\|, \Delta_{\rm g}] \tag{2.5}$$

for some constants $\gamma_{\rm g}$ and $\Delta_{\rm g} > 0$. Then clearly we must have that $\lim_{k\to\infty} g_k = 0$ provided f is bounded below on the level set $\mathcal{L}(x_0)$ where $\mathcal{L}(x) \stackrel{\rm def}{=} \{z \in {\rm I\!R}^n : f(z) \leq f(x)\}$.

The requirement (2.5) is reasonable for monotonic trust-region methods under standard assumptions. In particular, suppose that we apply the "basic-trust-region" algorithm (BTR) [2, Alg.6.1.1] to minimize $f_k^{\rm S}(y)$, starting from $y_{k,0} = 0$ with trust-region radius $\Delta_{k,0} \geq \Delta_{\rm g} > 0$ for all $k \geq 0$, generating the approximations $\{y_{k,\ell}\}_{\ell\geq 0}$ and radii $\{\Delta_{k,\ell}\}_{\ell\geq 0}$ and terminating at iteration e with some $y_k \stackrel{\rm def}{=} y_{k,e} \neq 0$. Furthermore suppose that $g_{k,\ell}^{\rm S} = g_k^{\rm S}(y_{k,\ell})$, that the model $m_{k,\ell}^{\rm S}(w) \stackrel{\rm def}{=} w^T g_{k,\ell}^{\rm S} + \frac{1}{2} w^T B_{k,\ell}^{\rm S} w$, and that we assume that both the true Hessian $H_k^{\rm S}(0)$ and those of models $B_{k,\ell}^{\rm S}$ employed at $y = 0$ remain bounded. Then standard analysis [2, Cor.6.3.2] shows that the first "successful" iterate $y_{k,s+1} \neq 0$ generated will be such that

$$\begin{aligned} f(x_k) - f(x_{k+1}) = & \;\; f_k^{\rm S}(0) - f_k^{\rm S}(y_{k,e}) \\ \geq & \;\; f_k^{\rm S}(y_{k,0}) - f_k^{\rm S}(y_{k,s+1}) \\ \geq & \;\; \eta_1 [m_{k,s}^{\rm S}(0) - m_{k,s}^{\rm S}(y_{k,s+1})] \\ \geq & \;\; \tfrac{1}{2}\eta_1 \|g_{k,s}^{\rm S}\| \min\left[\frac{\|g_{k,s}^{\rm S}\|}{1 + \|B_{k,s}^{\rm S}\|}, \Delta_{k,s}\right], \end{aligned} \tag{2.6}$$

where the value $\eta_1 \in (0,1)$ controls the acceptance of the step. But by assumption $1 + \|B_{k,s}^{\rm S}\| \leq \kappa_{\rm umh}$ and $\|H_k^{\rm S}(0)\| \leq \kappa_{\rm ufh}$ for appropriate constants, while it can easily be shown that

$$\Delta_{k,s} \geq \min\left[\Delta_{k,0}, \|g_{k,s}^{\rm S}\| \frac{\gamma_1(1-\eta_1)}{\kappa_{\rm umh} + \kappa_{\rm ufh}}\right], \tag{2.7}$$

where γ_1 is the trust-region contraction factor in the unsuccessful case [2, essentially Thm. 6.4.2]. Moreover,

$$g_{k,s}^{\rm S} = g_{k,0}^{\rm S} = g_k^{\rm S}(0) = S_k^T g_k$$

and thus by construction of S_k and (2.4)

$$\|g^{\rm S}_{k,s}\| = \|S_k^T g_k\| \geq |s_{k,1}^T g_k| \geq \kappa_{\rm g}\|g_k\|. \tag{2.8}$$

The required bound (2.5) then follows from (2.6)–(2.8). Thus if the subspace spanned by the columns of S_k contains g_k, or has a substantial component of g_k in the sense of (2.4), the use of BTR ensures global convergence to a first-order critical point of bounded f.

Now let λ_k be the left-most eigenvalue of H_k, and suppose that one or more of the columns of S_k—say the second, $s_{k,2}$, although in principle it might be the same vector $s_{k,1}$ which gave sufficient descent in (2.4)—provides a significant component of the corresponding extreme eigenvector in the sense that

$$\lambda_k \leq s_{k,2}^T H_k s_{k,2} \leq \kappa_{\rm H}\lambda_k \tag{2.9}$$

for some $\kappa_{\rm H} > 0$. Furthermore, suppose that, in addition to (2.5), the new iterate x_{k+1} improves f in as much as

$$f(x_{k+1}) \leq f(x_k) + \gamma_{\rm H}\lambda_k \min[\lambda_k^2, \Delta_{\rm H}] \tag{2.10}$$

for some constant $\gamma_{\rm H}$ and $\Delta_{\rm H} > 0$. Then clearly we must have that $\lim_{k\to\infty}\lambda_k \geq 0$ provided f is once again bounded below on the level set $\mathcal{L}(x_0)$.

Fortunately, BTR may be extended to ensure (2.10). In particular, suppose that we now insist that BTR uses the second-order model $m^{\rm SS}_{k,\ell}(w) \stackrel{\rm def}{=} w^T g^{\rm S}_{k,\ell} + \frac{1}{2}w^T H^{\rm S}_k(0)w$ when employed at $y = 0$ whenever $H^{\rm S}_k(0)$ is indefinite (actually this is only required when $\|g_k\|$ is small), that $H(x)$ is Lipschitz continuous with Lipschitz constant $\kappa_{\rm lch}$ within $\mathcal{L}(x_0)$, and that in addition to (2.6) we require that

$$m^{\rm SS}_{k,s}(0) - m^{\rm SS}_{k,s}(y_{k,i+1}) \geq -\tfrac{1}{2}\kappa_{\rm H}\lambda_k\Delta_{k,i}^2; \tag{2.11}$$

the reduction (2.11) can be guaranteed along the second component of y because of (2.9) [2, Thm.6.6.1]. Thus the first "successful" iterate $y_{k,s+1} \neq 0$ generated for which (2.11) is also required will satisfy

$$f(x_k) - f(x_{k+1}) \geq f^{\rm S}_k(y_{k,0}) - f^{\rm S}_k(y_{k,s+1}) \geq -\tfrac{1}{2}\eta_1\kappa_{\rm H}\lambda_k\Delta_{k,s}^2. \tag{2.12}$$

But since it follows immediately from Taylor's theorem and Lipschitz continuity that $|f^{\rm S}_k(y_{k,i+1}) - m^{\rm SS}_{k,s}(y_{k,i+1})| \leq \frac{1}{2}\kappa_{\rm lch}\Delta_{k,i}^3$, it is straightforward to show from (2.11) that the trust region radius satisfies

$$\Delta_{k,s} \geq \min\left[\Delta_{k,0}, |\lambda_k|\frac{\kappa_{\rm H}\gamma_1(1-\eta_1)}{\kappa_{\rm lch}}\right] \tag{2.13}$$

as such a radius will ensure that $y_{k,s+1}$ is successful. The required reduction (2.10) then follows directly from (2.12) and (2.13).

3 Subspace surrogates

Now suppose that $P \in \mathbb{R}^{n \times n_p}$ is a fixed subspace-prolongation matrix, and that

$$S_k = \begin{pmatrix} P & \frac{g_k}{\|g_k\|} \end{pmatrix} \in \mathbb{R}^{n \times n_p + 1}.$$

Suppose in addition that there is a suitable *surrogate* $f^{\mathrm{P}}(v)$ of $f(Pv)$, that is to say an approximation of the objective function $f(x)$ resptricted to the subspace $x = Pv$ for given $v \in \mathbb{R}^{n_p}$. Our aim is to approximately minimize $f_k^{\mathrm{S}}(y)$, where $y = (v, \mu)$. Rather than building a quadratic model of $f_k^{\mathrm{S}}(u+s, \nu+\sigma)$ as is standard in trust-region methods, it may be more appropriate instead to consider the non-quadratic approximation

$$\begin{array}{rl} m(s,\sigma) = & f^{\mathrm{P}}(v+s) - f^{\mathrm{P}}(u) + f_k^{\mathrm{S}}(v,\nu) + s^T P^T g(x) - s^T \nabla f^{\mathrm{P}}(u) \\ & + \sigma ||g(x)|| + \frac{1}{2}\sigma^2 g(x)^T B g(x) \end{array},$$

where B is a suitable symmetric approximation to $H(x)$. The model $m(s,\sigma)$ has the desirable properties that $m(0,0) = f_k^{\mathrm{S}}(y)$ and $\nabla m(0,0) = \nabla f_k^{\mathrm{S}}(y)$, and thus approximate minimization of m within suitably adjusted trust regions leads to critical points of f_k^{S}. Note, however, that the Hessians of the function and model do not agree so fast asymptotic convergence is not assured.

4 An Enriched Recursive Multilevel Approach

Again consider the problem of minimizing the twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. Given an estimate x_k of the required solution we wish to compute a trial step s_k within a trust-region setting—let the trust region radius be Δ. We also assume that we have (for simplicity) two levels of granularity. Therefore, we suppose we are given $R : \mathbb{R}^n \to \mathbb{R}^{n_c}$ (the restriction operator) and $P : \mathbb{R}^{n_c} \to \mathbb{R}^n$ (the prolongation operator) such that $R = P^T$. Since we are only considering two levels, we can refer to them as the fine level and the course level. One method for generating a trial step is to compute the standard trust-region step given by

$$s_k = \arg\min_{s \in \mathbb{R}^n} g_k^T s + \tfrac{1}{2} s^T H_k s, \quad \text{subject to} \quad \|s\| \le \Delta \tag{4.14}$$

where g_k and H_k are the gradient and Hessian of f evaluated at x_k, respectively. However, another option is to compute a trial step from the problem

$$\min_{y \in \mathbb{R}^{n_c+1}} f(x_k + S_k y) \stackrel{\text{def}}{=} f_k^{\mathrm{S}}(y), \tag{4.15}$$

where $S_k = \begin{pmatrix} P & g_k \end{pmatrix}$. This problem may be solved by solving a sequence of trust region problems of the form

$$\min_{\Delta y \in \mathbb{R}^{n_c+1}} \Delta y^T g_{k,\ell}^{\mathrm{S}} + \tfrac{1}{2} \Delta y^T H_{k,\ell}^{\mathrm{S}} \Delta y, \quad \text{subject to} \quad \|\Delta y\| \le \Delta \tag{4.16}$$

where $g^{\rm S}_{k,\ell} = g^{\rm S}_k(y_{k,\ell})$ and $H^{\rm S}_{k,\ell} = H^{\rm S}_k(y_{k,\ell})$ are the gradient and Hessian of $f^{\rm S}_k$ evaluated at $y_{k,\ell}$, respectively. Once a solution y^* has been computed, then we may define the trial step as $s_k = S_k y^*_k$, so that the trial point is $x_k + S_k y^*_k$.

If the gradient at the course level is "small", then we may also try solving problem (4.14) in the spirit of [], which was motivated by multigrid for solving positive-definite systems of equations. Thus we could perform a cycle of coordinate searchs for problem (4.14), restrict to the course level, perform smoothing along S_k, and then prolongate back to the fine grid, to be followed again by smoothing.

This general idea could be applied recursively if more than two levels of coursening exist.

4.1 Details

Suppose that $m_0 \stackrel{\rm def}{=} n > m_1 > \ldots > m_i > m_{i+1} > \ldots > m_\ell$ are specified dimensions of the nested subspaces (levels), that $y_{k,i} \in \mathbb{R}^{m_i}$ for $1 \leq i < \ell$ are given vectors of values, and that we define independent nested subspace vectors via

$$x = x_k + S_{k,1} y_1 \;\text{ and }\; y_i = y_{k,i} + S_{k,i+1} y_{i+1} \;\text{ for } 1 \leq i < \ell, \tag{4.17}$$

where $S_{k,i} \in \mathbb{R}^{m_{i-1} \times m_i}$, $1 \leq i < \ell$, are given subspace matrices. Now define

$$f_{k,1}(y_1) = f(x_k + S_{k,1} y_1) \;\text{ and }\; f_{k,i+1}(y_{i+1}) = f_{k,i}(y_{k,i} + S_{k,i+1} y_{i+1}) \;\text{ for } 1 \leq i < \ell.$$

Then it follows immediately that

$$\begin{aligned} \nabla_{y_1} f_{k,1}(y_1) &= S^T_{k,1} \nabla_x f(x_k + S_{k,1} y_1) \equiv S^T_{k,1} g(x_k + S_{k,1} y_1) \\ \text{and }\; \nabla_{y_{i+1}} f_{k,i+1}(y_{i+1}) &= S^T_{k,i+1} \nabla_{y_i} f_{k,i}(y_{k,i} + S_{k,i+1} y_{i+1}) \;\text{ for } 1 \leq i < \ell \end{aligned} \tag{4.18}$$

and that

$$\begin{aligned} \nabla_{y_1 y_1} f_{k,1}(y_1) &= S^T_{k,1} \nabla_{xx} f(x_k + S_{k,1} y_0) S_{k,1} \equiv S^T_{k,1} H(x_k + S_{k,1} y_0) S_{k,1} \\ \text{and }\; \nabla_{y_{i+1} y_{i+1}} f_{k,i+1}(y_{i+1}) &= S^T_{k,i+1} \nabla_{y_i y_i} f_{k,i}(y_{k,i} + S_{k,i+1} y_{i+1}) S_{k,i+1} \;\text{ for } 1 \leq i < \ell. \end{aligned} \tag{4.19}$$

In particular, if we specify $y_\ell = \bar{y}_{k,\ell}$ for given $\bar{y}_{k,\ell}$ and recover the corresponding $y_i = \bar{y}_{k,i}$ and $x = \overline{x}_k$ recursively from (4.17) via

$$\bar{y}_{k,i} = y_{k,i} + S_{k,i+1} \bar{y}_{k,i+1}, \;\text{ for } \ell > i \geq 1, \text{ and } \; \overline{x}_k = x_k + S_{k,1} \bar{y}_{k,1}, \tag{4.20}$$

and if we define

$$\bar{g}_{k,0} \equiv \bar{g}_k \stackrel{\rm def}{=} g(\bar{x}_k),\; \bar{H}_{k,0} \stackrel{\rm def}{=} H(\bar{x}_k),\; \overline{g}_{k,i} \stackrel{\rm def}{=} \nabla_{y_i} f_{k,i}(\overline{y}_{k,i}) \;\text{ and }\; \overline{H}_{k,i} \stackrel{\rm def}{=} \nabla_{y_i y_i} f_{k,i}(\overline{y}_{k,i}),$$

for $1 \leq i \leq \ell$, gradients and Hessians for all levels may be obtained recursively from (4.18) and (4.19) as

$$\overline{g}_{k,i} = S^T_{k,i} \overline{g}_{k,i-1} \;\text{ and }\; \overline{H}_{k,i} = S^T_{k,i} \overline{H}_{k,i-1} S_{k,i}$$

for $1 \leq i \leq \ell$.

We now consider particular subspace matrices constructed by augmenting standard multigrid prolongation matrices at each level with the corresponding gradients g_k (level 0) and $g_{k,i} \stackrel{\text{def}}{=} \nabla_{y_i} f_{k,i}(0)$ for $1 \leq i \leq \ell$ (levels 1 to ℓ). Thus if we define $n_0 \stackrel{\text{def}}{=} n$ and $n_i \stackrel{\text{def}}{=} m_i - 1$ for $1 \leq i \leq \ell$, and suppose that $P_i \in \mathbb{R}^{n_i \times n_{i+1}}$ are given prolongation matrices for $0 \leq i < \ell$, we consider the subspace matrices

$$S_{k,1} = \begin{pmatrix} P_0 & \dfrac{g_k}{\|g_k\|} \end{pmatrix} \in \mathbb{R}^{n \times n_1 + 1} \quad \text{and}$$
$$S_{k,i+1} = \begin{pmatrix} \begin{pmatrix} P_i \\ 0 \end{pmatrix} & \dfrac{g_{k,i}}{\|g_{k,i}\|} \end{pmatrix} \in \mathbb{R}^{n_i + 1 \times n_{i+1} + 1} \quad \text{for } 1 \leq i < \ell.$$

Thus

$$\overline{g}_{k,i+1} = S_{k,i+1}^T \overline{g}_{k,i} = \begin{pmatrix} \begin{pmatrix} P_i^T & 0 \end{pmatrix} \\ \dfrac{g_{k,i}^T}{\|g_{k,i}\|} \end{pmatrix} \overline{g}_{k,i} = \begin{pmatrix} \begin{pmatrix} P_i^T & 0 \end{pmatrix} \overline{g}_{k,i} \\ \dfrac{\overline{g}_{k,i}^T g_{k,i}}{\|g_{k,i}\|} \end{pmatrix}$$

while

$$g_{k,i+1} = S_{k,i+1}^T g_{k,i} = \begin{pmatrix} \begin{pmatrix} P_i^T & 0 \end{pmatrix} \\ \dfrac{g_{k,i}^T}{\|g_{k,i}\|} \end{pmatrix} g_{k,i} = \begin{pmatrix} \begin{pmatrix} P_i^T & 0 \end{pmatrix} g_{k,i} \\ \|g_{k,i}\| \end{pmatrix}.$$

References

[1] A. R. Conn, N. I. M. Gould, A. Sartenaer, and Ph. L. Toint. On iterated-subspace minimization methods for nonlinear optimization. In L. Adams and L. Nazareth, editors, *Proceedings on Linear and Nonlinear Conjugate Gradient-Related Methods*, pages 50–78, Philadelphia, 1996. SIAM.

[2] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. SIAM, Philadelphia, 2000.

[3] H. S. Dollar, N. I. M. Gould, and D. P. Robinson. On solving trust-region and other regularised subproblems in optimization. Technical Report RAL-TR-2009-003, Rutherford Appleton Laboratory, 2009.

[4] M. Elad, B. Matalonb, and M. Zibulevsky. Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization. *Applied and Computational Harmonic Analysis*, 23(3):346–367, 2007.

[5] J. B. Erway and P. E. Gill. A subspace minimization method for the trust-region step. Technical Report NA 08-1, Department of Mathematics, University of California, San Diego, 2008.

[6] J. B. Erway, P. E. Gill, and J. D. Griffin. Iterative methods for finding a trust-region step. Technical Report NA 07-2, Department of Mathematics, University of California, San Diego, 2007.

[7] D. M. Gay. Computing optimal locally constrained steps. *SIAM Journal on Scientific and Statistical Computing*, 2(2):186–197, 1981.

[8] N. I. M. Gould, S. Lucidi, M. Roma, and Ph. L. Toint. Solving the trust-region subproblem using the Lanczos method. *SIAM Journal on Optimization*, 9(2):504–525, 1999.

[9] W. W. Hager. Minimizing a quadratic over a sphere. *SIAM Journal on Optimization*, 12(1):188–208, 2001.

[10] J. J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 4(3):553–572, 1983.

[11] G. Narkiss and M. Zibulevsky. Sequential subspace optimization method for large-scale unconstrained problems. Technical Report CCIT 559, Electrical Engineering Department, Technion, Haifa, Israel, 2005.

[12] Y. Yuan. Subspace techniques for nonlinear optimization. In R. Jeltsch, D.Q. Li, and I. H. Sloan, editors, *Some Topics in Industrial and Applied Mathematics*, Series in Contemporary Applied Mathematics CAM 8, pages 206–218, Beijing, 2007. Higher Education Press.

[13] Y. Yuan. Subspace methods for large scale nonlinear equations and nonlinear least squares. *Optimizaiton and Engineering*, (to appear), 2009.

[14] M. Zibulevsky. SESOP-TN: combining sequential subspace optimization with truncated Newton method. Technical report, Computer Science Department, Technion, Haifa, Israel, 2008.