PANM 13

Ladislav Lukšan; Ctirad Matonoha; Jan Vlček

Interior-point method for large-scale l_1 optimization

In: Jan Chleboun and Karel Segeth and Tomáš Vejchodský (eds.): Programs and Algorithms of Numerical Mathematics, Proceedings of Seminar. Prague, May 28-31, 2006. Institute of Mathematics AS CR, Prague, 2006. pp. 190–195.

Persistent URL: http://dml.cz/dmlcz/702836

INTERIOR-POINT METHOD FOR LARGE-SCALE l_1 OPTIMIZATION*

Ladislav Lukšan, Ctirad Matonoha, Jan Vlček

Consider the l_1 optimization problem: Minimize function

$$F(x) = \sum_{i=1}^{m} |f_i(x)|, \tag{1}$$

where $f_i : R^n \to R$, $0 \le i \le m$ (m is usually large), are smooth functions depending on a small number of variables. We will assume that these functions are twice continuously differentiable with bounded first and second-order derivatives in a sufficiently large region $\mathcal{D}$.

Minimization of F is equivalent to the sparse nonlinear programming problem with $n+m$ variables $x \in R^n$, $z \in R^m$:

$$\text{minimize} \quad \sum_{i=1}^{m} z_i \quad \text{subject to} \quad -z_i \le f_i(x) \le z_i, \quad 1 \le i \le m. \tag{2}$$

In this contribution, we introduce a trust-region interior-point method for nonconvex nonlinear programming that utilizes a special structure of problem (2). All theoretical results are given without proofs. These proofs can be found in [5].

The constrained problem (2) is replaced by a sequence of unconstrained problems

$$\begin{aligned} \text{minimize} \quad B(x, z; \mu) \quad &= \quad \sum_{i=1}^{m} z_i - \mu \sum_{i=1}^{m} \log(z_i - f_i(x)) - \mu \sum_{i=1}^{m} \log(z_i + f_i(x)) \\ &= \quad \sum_{i=1}^{m} z_i - \mu \sum_{i=1}^{m} \log(z_i^2 - f_i^2(x)), \end{aligned} \tag{3}$$

where $z_i > |f_i(x)|$, $1 \le i \le m$, and $\mu > 0$ (we assume that $\mu \to 0$ monotonically). Barrier function (3) remains unchanged if we replace problem (2) by equivalent problem

$$\text{minimize} \quad \sum_{i=1}^{m} z_i \quad \text{subject to} \quad f_i^2(x) \le z_i^2, \quad 1 \le i \le m. \tag{4}$$

The necessary first-order (KKT) conditions for the solution of (4) have the form

$$\sum_{i=1}^{m} 2w_i f_i(x) \nabla f_i(x) = 0, \quad 2w_i z_i = 1, \quad w_i \ge 0, \quad w_i(z_i^2 - f_i^2(x)) = 0, \quad 1 \le i \le m, \tag{5}$$

*This work was supported by the Grant Agency of the Czech Academy of Sciences, project No. IAA1030405 and the institutional research plan No. AV0Z10300504

where w_i, $1 \le i \le m$, are Lagrange multipliers. Since $z_i = |f_i(x)|$, $1 \le i \le m$, at the solution of (4), we can write (5) in a simpler equivalent form

$$\sum_{i=1}^{m} u_i \nabla f_i(x) = 0, \quad \frac{u_i z_i}{f_i(x)} = 1, \quad z_i^2 - f_i^2(x) = 0, \quad 1 \le i \le m, \tag{6}$$

where $u_i = 2 w_i f_i(x)$ for $1 \le i \le m$.

The special structure of problem (3) allows us to obtain minimizer $z(x;\mu) \in R^m$ of function $B(x, z; \mu)$ for a given $x \in R^n$.

Lemma 1. *Function $B(x, z; \mu)$ (with x fixed) has the unique stationary point, which is its global minimizer. This stationary point is characterized by equations*

$$\frac{2\mu z_i(x;\mu)}{z_i^2(x;\mu) - f_i^2(x)} = 1 \quad \text{or} \quad z_i^2(x;\mu) - f_i^2(x) = 2\mu z_i(x;\mu), \quad 1 \le i \le m, \tag{7}$$

which have solutions

$$z_i(x;\mu) = \mu + \sqrt{\mu^2 + f_i^2(x)}, \quad 1 \le i \le m. \tag{8}$$

Assuming $z = z(x;\mu)$, we denote

$$B(x;\mu) = \sum_{i=1}^{m} z_i(x;\mu) - \mu \sum_{i=1}^{m} \log(z_i^2(x;\mu) - f_i^2(x)) \tag{9}$$

and $u(x;\mu) = u(x, z(x;\mu); \mu)$. In this case, barrier function $B(x;\mu)$ depends only on x.

Lemma 2. *Consider barrier function (9). Then*

$$\nabla B(x;\mu) = g(x;\mu), \tag{10}$$

where $g(x;\mu) = A(x)u(x;\mu) = \sum_{i=1}^{m} \nabla f_i(x) u_i(x;\mu)$ with

$$u_i(x;\mu) = \frac{2\mu f_i(x)}{z_i^2 - f_i^2(x)}, \quad 1 \le i \le m, \tag{11}$$

and

$$\nabla^2 B(x;\mu) = G(x;\mu) + A(x)V(x;\mu)A^T(x), \tag{12}$$

where

$$G(x;\mu) = \sum_{i=1}^{m} u_i(x;\mu) G_i(x) \tag{13}$$

with $G_i(x) = \nabla^2 f_i(x)$, $1 \le i \le m$, and $V(x;\mu) = \operatorname{diag}(v_1(x;\mu), \ldots, v_m(x;\mu))$ with

$$v_i(x;\mu) = \frac{2\mu}{z_i^2(x;\mu) + f_i^2(x)}, \quad 1 \le i \le m. \tag{14}$$

Lemma 3. *Let $\nabla^2 B(x;\mu)d = -\nabla B(x;\mu)$. If matrix $G(x;\mu)$ is positive definite, then $d^T g(x;\mu) < 0$ (direction vector d is descent for $B(x;\mu)$).*

Since positive definiteness of matrix $G(x;\mu)$ is not assured, the standard line-search methods cannot be used. For this reason, trust-region methods were developed. These methods use the direction vector obtained as an approximate minimizer of the quadratic subproblem

$$\text{minimize} \quad Q(d) = \frac{1}{2} d^T \nabla^2 B(x;\mu) d + g^T(x;\mu) d \quad \text{subject to} \quad \|d\| \le \Delta, \tag{15}$$

where Δ is the trust region radius. Direction vector d serves for obtaining new point $x^+ \in R^n$. Denoting

$$\rho(d) = \frac{B(x+d;\mu) - B(x;\mu)}{Q(d)}, \tag{16}$$

we set

$$x^+ = x \quad \text{if} \quad \rho(d) \le 0, \quad \text{or} \quad x^+ = x + d \quad \text{if} \quad \rho(d) > 0. \tag{17}$$

Finally, we update the trust region radius in such a way that

$$\begin{aligned} \Delta^+ &= \underline{\beta}\Delta \quad \text{if} \quad \rho(d) < \underline{\rho}, \\ \Delta^+ &= \Delta \quad \text{if} \quad \underline{\rho} \le \rho(d) \le \overline{\rho}, \\ \Delta^+ &= \overline{\beta}\Delta \quad \text{if} \quad \overline{\rho} < \rho(d), \end{aligned} \tag{18}$$

where $0 < \underline{\rho} < \overline{\rho} < 1$ and $0 < \underline{\beta} < 1 < \overline{\beta}$.

Now we are in a position to describe the basic algorithm.

Algorithm 1.

Data: Termination parameter $\underline{\varepsilon} > 0$, minimum value of the barrier parameter $\underline{\mu} > 0$, rate of the barrier parameter decrease $0 < \tau < 1$, trust-region parameters $0 < \underline{\rho} < \overline{\rho} < 1$, trust-region coefficients $0 < \underline{\beta} < 1 < \overline{\beta}$, step bound $\overline{\Delta} > 0$.

Input: Sparsity pattern of matrix A. Initial estimation of vector x.

Step 1: *Initiation.* Choose initial barrier parameter $\mu > 0$ and initial trust-region radius $0 < \Delta \le \overline{\Delta}$. Determine the sparsity pattern of matrix $\nabla^2 B$ from the sparsity pattern of matrix A. Carry out symbolic decomposition of $\nabla^2 B$. Compute values $f_i(x)$, $1 \le i \le m$, and $F(x) = \sum_{1 \le i \le m} |f_i(x)|$. Set $k := 0$ (iteration count).

Step 2: *Termination.* Determine vector $z(x;\mu)$ by (8) and vector $u(x;\mu)$ by (11). Compute matrix $A(x)$ and vector $g(x;\mu) = A(x)u(x;\mu)$. If $\mu \le \underline{\mu}$ and $\|g(x;\mu)\| \le \underline{\varepsilon}$, then terminate the computation. Otherwise set $k := k+1$.

Step 3: *Approximation of the Hessian matrix.* Compute approximation of matrix $G(x;\mu)$ by using differences $A(x+\delta v)u(x;\mu) - g(x;\mu)$ for a suitable set of vectors v (see [1]). Determine Hessian matrix $\nabla^2 B(x;\mu)$ by (12).

Step 4: *Direction determination.* Determine vector d as an approximate solution of trust-region subproblem (15).

Step 5: *Step-length selection.* Determine x^+ by (17) and set $x := x^+$. Compute values $f_i(x)$, $1 \leq i \leq m$, and $F(x) = \sum_{1 \leq i \leq m} |f_i(x)|$.

Step 6: *Trust-region update.* Determine new trust-region radius Δ by (18) and set $\Delta := \min(\Delta, \overline{\Delta})$.

Step 7: *Barrier parameter update.* If $\rho(d) \geq \underline{\rho}$ (where $\rho(d)$ is given by (16)), determine a new value of barrier parameter $\mu \geq \underline{\mu}$ (not greater than the current one) by the procedure described below. Go to Step 2.

The use of the maximum step-length $\overline{\Delta}$ has no theoretical significance but is very useful for practical computations. First, the problem functions can sometimes be evaluated only in a relatively small region (if they contain exponentials) so that the maximum step-length is necessary. Secondly, the problem can be very ill-conditioned far from the solution point, thus large steps are unsuitable. Finally, if the problem has more local solutions, a suitably chosen maximum step-length can cause a local solution with a lower value of F to be reached. Therefore, maximum step-length $\overline{\Delta}$ is a parameter, which is most frequently tuned.

Direction vector $d \in R^n$ should be a solution of the quadratic subproblem (15). Usually, an inexact approximate solution suffices. The dog-leg method described in [6], [2], seeks d as a linear combination of the Cauchy step $d_C = -(g^T g / g^T \nabla^2 B g) g$ and the Newton step $d_N = -(\nabla^2 B)^{-1} g$. The Newton step is computed by using either sparse Gill-Murray decomposition [4] or sparse Bunch-Parlett decomposition [3]. The sparse Gill-Murray decomposition has the form $\nabla^2 B + E = LDL^T = R^T R$, where E is a positive semidefinite diagonal matrix (which is equal to zero when $\nabla^2 B$ is positive definite), L is a lower triangular matrix, D is a positive definite diagonal matrix and R is an upper triangular matrix. The sparse Bunch-Parlett decomposition has the form $\nabla^2 B = PLML^T P^T$, where P is a permutation matrix, L is a lower triangular matrix and M is a block-diagonal matrix with 1×1 or 2×2 blocks (which is indefinite when $\nabla^2 B$ is indefinite). The following algorithm is a typical implementation of the dog-leg method.

Algorithm A: Data $\Delta > 0$.

Step 1: If $g^T \nabla^2 B g \leq 0$, set $s := -(\Delta / \|g\|) g$ and terminate the computation.

Step 2: Compute the Cauchy step $d_C = -(g^T g / g^T \nabla^2 B g) g$. If $\|d_C\| \geq \Delta$, set $d := (\Delta / \|d_C\|) d_C$ and terminate the computation.

Step 3: Compute the Newton step $d_N = -(\nabla^2 B)^{-1} g$. If $(d_N - d_C)^T d_C \geq 0$ and $\|d_N\| \leq \Delta$, set $d := d_N$ and terminate the computation.

Step 4: If $(d_N - d_C)^T d_C \geq 0$ and $\|d_N\| > \Delta$, determine number θ in such a way that $d_C^T d_C / d_C^T d_N \leq \theta \leq 1$, choose $\alpha > 0$ such that $\|d_C + \alpha(\theta d_N - d_C)\| = \Delta$, set $d := d_C + \alpha(\theta d_N - d_C)$ and terminate the computation.

Step 5: If $(d_N - d_C)^T d_C < 0$, choose $\alpha > 0$ such that $\|d_C + \alpha(d_C - d_N)\| = \Delta$, set $d := d_C + \alpha(d_C - d_N)$ and terminate the computation.

The above algorithm generates direction vectors such that

$$\begin{aligned} \|d\| &\leq \Delta, \\ \|d\| &< \Delta \quad \Rightarrow \quad \nabla^2 B d = -g, \\ -Q(d) &\geq \underline{\sigma}\|g\| \min\left(\Delta, \frac{\|g\|}{\|\nabla^2 B\|}\right), \end{aligned}$$

where $0 < \underline{\sigma} < 1$ is a constant. These inequalities imply (see [7]), that a constant $0 < \underline{c} < 1$ exists such that

$$\|d\| \geq \underline{c}\gamma/\overline{B}, \tag{19}$$

where γ is the minimum norm of gradients that have been computed and $\overline{B}$ is an upper bound for $\|\nabla^2 B\|$ (assuming $\mu \geq \underline{\mu} > 0$, we can set $\overline{B} = m(\overline{G} + \overline{g}^2/(2\underline{\mu}))$). Thus

$$B(x+d;\mu) - B(x;\mu) \leq \underline{\rho} Q(d) \leq -\underline{\rho}\,\underline{\sigma}\,\underline{c}\frac{\gamma^2}{\overline{B}} \quad \text{if} \quad \rho \geq \underline{\rho} \tag{20}$$

by (17) and (19).

Algorithm 1 is sensitive on the way in which the barrier parameter decreases. We have tested various possibilities for the barrier parameter update including simple geometric sequences, which were proved to be unsuitable. Better results were obtained by setting

$$\mu_{k+1} = \mu_k \quad \text{if} \quad \|g_k\|^2 > \tau\mu_k \quad \text{or} \quad \mu_{k+1} = \max(\underline{\mu}, \|g_k\|^2) \quad \text{if} \quad \|g_k\|^2 \leq \tau\mu_k, \tag{21}$$

where $0 < \tau < 1$.

In the subsequent considerations, we will assume that $\underline{\varepsilon} = \underline{\mu} = 0$ and all computations are exact.

Lemma 4. *Let Assumption 3 be satisfied. Then values $\{\mu_k\}_1^\infty$, generated by Algorithm 1, form a non-increasing sequence such that $\mu_k \to 0$.*

Lemma 5. *The inequality*

$$B(x_{k+1};\mu_{k+1}) \leq B(x_{k+1};\mu_k) - L(\mu_{k+1} - \mu_k) \tag{22}$$

holds with some $L \in R$.

Theorem 1. *Consider sequence $\{x_k\}_1^\infty$, generated by Algorithm 1. Then*

$$\liminf_{k\to\infty} \sum_{i=1}^m u_i(x_k;\mu_k) g_i(x_k) = 0$$

and

$$u_i(x_k;\mu_k) = \frac{f_i(x_k)}{z_i(x_k;\mu_k)}, \quad \lim_{k\to\infty}(z_i^2(x_k;\mu_k) - f_i^2(x_k)) = 0$$

for $1 \leq i \leq m$.

Remark 1. If we replace (17) by

$$x^+ = x \quad \text{if} \quad \rho(d) < \underline{\rho}, \quad \text{or} \quad x^+ = x + d \quad \text{if} \quad \rho(d) \geq \underline{\rho} \tag{23}$$

in Algorithm 1, then $\lim_{k\to\infty} \|g(x_k; \mu_k)\| = 0$.

Corollary 1. *Let assumptions of Theorem 1 and (23) hold. Then every cluster point $x \in R^n$ of sequence $\{x_k\}_1^\infty$ satisfies KKT conditions (6), where $u \in R^m$ is a cluster point of sequence $\{u(x_k; \mu_k)\}_1^\infty$.*

The efficiency of Algorithm 1 was tested by using extensive collections of test problems. The results are given in [5].

References

[1] T.F. Coleman, J.J. Moré: *Estimation of sparse Hessian matrices and graph coloring problems.* Mathematical Programming **28**, 1984, 243–270.

[2] J.E. Dennis, H.H.W. Mei: *An unconstrained optimization algorithm which uses function and gradient values.* Report No. TR 75-246, 1975.

[3] I.S. Duff, M. Munksgaard, H.B. Nielsen, J.K. Reid: *Direct solution of sets of linear equations whose matrix is sparse and indefinite.* J. Inst. Maths. Applics. **23**, 1979, 235–250.

[4] P.E. Gill, W. Murray: *Newton type methods for unconstrained and linearly constrained optimization.* Mathematical Programming **7**, 1974, 311–350.

[5] L. Lukšan, C. Matonoha, J. Vlček: Trust-region interior-point method for large sparse l_1 optimization. Report V-942, Prague, ICS AS CR, 2005 (to appear in Optimization Methods and Software).

[6] M.J.D. Powell: *A new algorithm for unconstrained optimization.* In: Nonlinear Programming, J.B. Rosen, O.L. Mangasarian, K. Ritter (eds.), Academic Press, London 1970.

[7] M.J.D. Powell: *On the global convergence of trust region algorithms for unconstrained optimization.* Mathematical Programming **29**, 1984, 297–303.