Murphy, Frederic H.

Working Paper

Improved Convergence Rate Results for a Class of Exponential Penalty Functions

Discussion Paper, No. 115

Provided in Cooperation with:
Kellogg School of Management - Center for Mathematical Studies in Economics and Management Science, Northwestern University

Suggested Citation: Murphy, Frederic H. (1974) : Improved Convergence Rate Results for a Class of Exponential Penalty Functions, Discussion Paper, No. 115, Northwestern University, Kellogg School of Management, Center for Mathematical Studies in Economics and Management Science, Evanston, IL

This Version is available at:
https://hdl.handle.net/10419/220475

Discussion Paper No. 115

Improved Convergence Rate Results for a Class of Exponential Penalty Functions

by

Frederic H. Murphy

November 1974

ABSTRACT

An improved theoretical rate of convergence is shown for a member of the class of exponential penalty function algorithms. We show that the algorithm has the structure of an asymptotic Newton algorithm, which means this algorithm has a superlinear convergence rate.

In [9] we presented an exponential penalty function algorithm for the following nonlinear programming NLP:

$$\underset{x\in E^n}{\text{minimize}}\ f(x) \tag{1}$$

$$\text{subject to}\quad g_i(x) \le 0 \qquad \text{for } i=1,\ldots,m \tag{2}$$

with an optimal solution x^*. The penalty function is

$$F_k(x) = f(x) - \sum_{i=1}^{m} \frac{1}{s_k} \exp[r_k g_i(x)] \tag{3}$$

where $r_k \ge s_k \ge 1$ and $r_k \to \infty$. We consider a variant of (3) where the rate of convergence is that of an asymptotic Newton algorithm as described in Mangasarian [8]. This variant meets one of the important concerns with asymptotic Newton algorithms, that is, it is everywhere second order differentiable if $f(\cdot)$, and $g_1(\cdot),\ldots,g_n(\cdot)$ are. Our algorithm is

Step 1: Minimize

$$f(x) + \sum_{i=1}^{m} \frac{\lambda_i^k}{r_k} \exp[r_k g_i(x)] \tag{4}$$

with the minimum at x_k.

Step 2: Set

$$\gamma_i^k = \lambda_i^k \exp[r_k g_i(x_k)], \tag{5}$$

then let

$$\lambda_i^{k+1} = \min[U_k, \max[L_k, \gamma_i^k]] \tag{6}$$

where $U_k > L_k > 0$, $U_k \to \infty$ and $L_k \to 0$, and $r_k L_k \to \infty$, and $U_k/r_k \to 0$ as $k \to \infty$.

Since this penalty function is clearly a member of the class of exponential penalty functions as described by Evans and Gould [3] we need not prove convergence here. To establish our convergence rate result we make the following assumptions:

1. NLP is a convex programming problem.
2. Slater's constraint qualification [7] is satisfied.
3. Either f(x) or one of the binding constraints is strictly convex.
4. The feasible region is compact.
5. Strict complementarity holds, that is, $g_i(x^*) = 0$ implies $\lambda_i^* > 0$ where λ_i^* is the optimal Lagrange multiplier for constraint i, i=1,...,m.
6. $\lambda_1^*,\ldots,\lambda_m^*$ are unique, x^* is a unique optimal solution.
7. $\{\nabla g_i(x^*) | g_i(x^*) = 0\ i=1,\ldots,m\}$ are linearly independent

In [10] it was shown that $\gamma_1^k,\ldots,\gamma_m^k$ remain uniformly bounded. As this is the standard proof as found in [4] it also is not repeated here.

Lemma 1. Let $I \subset \{1,\ldots,m\}$ be the index set of the binding constraints. With assumptions 1, 2, and 6 above,

$$\lim_{k\to\infty} r_k g_i(x_k) = 0 \tag{7}$$

for $i \in I$.

Proof. For $i \in I$, there is a K such that for $k \geq K$

$$\lambda_i^k = \min[U_t, \max[L_k, \gamma_i^k]] = \gamma_i^k . \tag{8}$$

Since $\gamma_i^k \to \lambda_i^*$ as $k \to \infty$ by the uniqueness of λ_i^*, $\lambda_i^k \to \lambda_i^*$ as $k \to \infty$. That is,

$$\lambda_i^{k+1} - \lambda_i^k = \lambda_i^k \exp[r_k g_i(x_k)] - \lambda_i^k$$

$$= \lambda_i^k [\exp[r_k g_i(x_k) - 1]]. \tag{9}$$

Since $\lambda_i^{k+1} - \lambda_i^k \to 0$ as $k \to \infty$ and $\lambda_i^* > 0$,

$$\exp[r_k g_i(x_k)] \to 1 \text{ as } k \to \infty \tag{10}$$

or $$r_k g_i(x_k) \to 0 \text{ as } k \to \infty .$$

Unlike the algorithms that fit into the classification of methods of multipliers [5], [8], [11], [12], with this algorithm we bear the computational burden of carrying along nonbinding constraint within the unconstrained minimization. To deal with this problem we have to analyze the behavior of the nonbinding constraints.

Lemma 2. There exists a K such that for $k \geq K$, and $i \notin I$, $\lambda_i^k = L_k$.

Proof. For $i \in I$, since $g_i(x^*) < 0$ there exists an $\varepsilon > 0$ and a K such that for $k \geq K$

$$g_i(x_k) \leq -\varepsilon. \tag{11}$$

That is

$$\exp[r_k g_i(x_k)] \leq \exp[-r_k \varepsilon]. \tag{12}$$

Consequently,

$$\gamma_i^k \leq \frac{\lambda_i^k}{r_k} e^{-r_k \varepsilon} \leq \frac{U_k}{r_k} e^{-r_k \varepsilon}. \tag{13}$$

Since $U_k/r_k \to 0$, $r_k \exp[-r_k \varepsilon] \to 0$ and $r_k L_k \to \infty$ as $k \to \infty$, there exists a K' such that $U_k/r_k \leq 1$ and $\exp[-r_k \varepsilon] \leq L_k$ for $k \geq K'$. Choose $k \geq \max[K,K']$ and the lemma holds.

Without loss of generality, let $I = \{1,\ldots,p\}$. We may then formulate a new nonlinear program NLP1:

$$\underset{x \in E^n}{\text{minimize}}\ f_o(r_k,x) = f(x) + \sum_{i=p+1}^{m} L_k/r_k \exp[r_k g_i(x)] \tag{14}$$

$$\text{subject to} \qquad g_i(x) = 0 \quad \text{for } i=1,\ldots,p. \tag{15}$$

Letting $\hat{x}_k$ be the optimal solution to NLP1 for a given r_k we see that $\hat{x}_k \to x^*$ as $k \to \infty$ and for k large enough $x_k = \hat{x}_k$; and for r_k large enough (14) is arbitrarily close to (1) over the feasible region defined by (2). We can now apply the duality results in Luenberger [6, p. 321]. We define the dual function φ of NLP1 for a given r_k as

$$\varphi(\lambda) = \min_{x \in E^n} \{ f(x) + \sum_{i=p+1}^{m} [L_k/r_k] \exp[r_k g_i(x)] + \sum_{i=1}^{p} [\lambda_i/r_k] [\exp[r_k g_i(x)]-1] \} \tag{16}$$

$$= \min_{x \in E^n} \{ f_o(r_k,x) + \sum_{i=1}^{p} [\lambda_i/r_k][\exp[r_k g_i(x)]-1] \}$$

where $\lambda = (\lambda_1,\ldots,\lambda_p)$.

The function φ is convex since we assumed $f(x)$, $g_1(x),\ldots,g_m(x)$ are convex.

Now by [6, p. 321] with $k \geq K$ as defined in Lemma 2

$$\nabla\varphi(\lambda^k) = \begin{bmatrix} 1/r_k [\exp[r_k g_1(x_k)]-1] \\ \vdots \\ 1/r_k [\exp[r_r g_p(x_k)]-1] \end{bmatrix} . \tag{17}$$

$$\text{Let } \nabla g(x) = \begin{bmatrix} \nabla g_1(x) \\ \vdots \\ \nabla g_p(x) \end{bmatrix} , \tag{18}$$

$$\text{Let } \Lambda(\lambda) = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \lambda_p \end{bmatrix} \tag{19}$$

$$\text{and let } H(r_k,x) = \begin{bmatrix} \exp[r_k g_1(x)] & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \exp[r_k g_p(x)] \end{bmatrix} . \tag{20}$$

and let $F_o(r_k,x)$ be the Hessian of $f_o(r_k,x)$ and let $G_i(x)$ be the Hessian of $g_i(x)$ for $i = 1,\ldots,p$.

From this we can construct the Hessian of $\varphi(\lambda^k)$:

$$H(r_k,x_k)\nabla g(x_k)[F_o(r_k,x_k) + \sum_{i=1}^{p} \lambda_i^k \exp[r_k g_i(x_k)]G_i(x_k) + r_k\nabla g(x_k)'A(\lambda^k)H(r_k,x_k)\nabla g_i(x_k)]^{-1} \nabla g_i(x_k)'H(r_k,x_k)' \tag{21}$$

The inverse of the matrix in the brackets exists since this matrix is positive definite. That comes about since we assumed strict convexity for one of the functions $f(x),g_1(x),\ldots,g_p(x)$ and all other terms in this matrix are positive semidefinite. Now (21) can be written as

$$M_k = H(r_k,x_k)\nabla g(x_k)[F_o(r_k,x_k) + \sum_{i=1}^{p} \lambda_i^k \exp[r_k g_i(x_k)]G_i(x_k) + r_k\nabla g(x_k)'A(\lambda^{k+1})\nabla g_i(x_k)]^{-1}\nabla g_i(x_k)'H(r_k,x_k)' \ . \tag{22}$$

To analyse (22) we now prove a simple extension of a lemma in Mangasarian [8].

Lemma 3. Let $C(\alpha) = B(A(\alpha) + B'Q(\alpha))^{-1}B'$ where B is a given $m \times n$ matrix of rank m, $A(\alpha)$ is an $n \times n$ matrix function on R, $Q(\alpha)$ is a differentiable $m \times m$ matrix function on R, $A(\alpha) + B'Q(\alpha)B$ is positive definite for $\alpha \geq \bar{\alpha}$ for some $\bar{\alpha}$, and for every $\varepsilon > 0$ there is an α_ε such that for $\alpha \geq \alpha_\varepsilon$ $|A(\alpha) - A| < \varepsilon$ where $A + B'Q(\alpha)B$ is positive definite for $\alpha \geq \bar{\alpha}$. Then there exists a constant matrix K such that for every $\delta > 0$ there is an α_δ with

$$|C(\alpha) - (Q(\alpha) + K)^{-1}| < \delta \tag{23}$$

for $\alpha \geq \alpha_\delta$.

Proof. Define $C_1(\alpha) = B(A + B'Q(\alpha))^{-1}B'$. By the continuity of the inverse of a matrix, for every δ there exists an ε such that

$$|C(\alpha) - C_1(\alpha)| < \delta \tag{24}$$

for $|A(\alpha)-A| < \varepsilon_\delta$. This means there exists an α_δ such that (24) holds for $\alpha \geq \alpha_\delta$.

The formula for differentiating the inverse of a matrix is $\frac{dC_1(\alpha)}{d\alpha} = - C_1(\alpha)^{-1} \frac{dC_1(\alpha)}{d\alpha} C_1(\alpha)^{-1}$. Differentiating $C_1(\alpha)$ we have

$$\begin{aligned} \frac{dC_1(\alpha)}{d\alpha} &= B\frac{d(A + B'Q(\alpha)B)^{-1}}{d\alpha} B' \\ &= -B^T(A + B'Q(\alpha)B)^{-1} \frac{d(A + B'Q(\alpha)B)}{d\alpha} (A + B'Q(\alpha)B)^{-1}B' \\ &= -C_1(\alpha) \frac{dQ(\alpha)}{d(\alpha)} C_1(\alpha) . \end{aligned} \tag{25}$$

Hence

$$\frac{dC(\alpha)^{-1}}{d\alpha} = \frac{dQ(\alpha)}{d(\alpha)} \text{ and } C_1(\alpha)^{-1} = Q(\alpha) + K . \tag{26}$$

That is,

$$C_1(\alpha) = (Q(\alpha) + K)^{-1}. \tag{27}$$

Substituting (27) into (24) we have (23).

For r_k sufficiently large there is a $\gamma < 0$ where $g_i(x_k) < \gamma$ for $i = 1,\ldots,p$. Letting A be the Hessian of $f(x)$ at x_k and $A(r_k)$ be the Hessian of $f_o(r_k,x)$ at x_k for $r_k = \alpha$, we know that $A(r_k) \to A$ as $r_k \to \infty$. Let $Q(r_k) = r_k\Lambda(\lambda^{k+1})$ and let $B = \nabla g(x_k)$. Remembering that $H(r_k,x_k) \to I$ as $r_k \to \infty$, we have the following theorem:

Theorem 1. If NLP satisfies 1,...,7, then the limit of $r_k M_k$ as $k \to \infty$ is $\Lambda(\lambda^*)^{-1}$.

Also, we know from (9) for i=1,...,p

$$\lambda_i^{k+1} - \lambda_i^k = \lambda_i^k[\exp[r_k g_i(x_k)]-1]. \tag{28}$$

Consequently,

$$\lambda^{k+1} = r_k\lambda^k\Lambda(\lambda^k)\nabla\varphi(\lambda^k). \tag{29}$$

Since $M_k \to \Lambda^{-1}(\lambda^k)$, we see that for large r $\Lambda(\lambda^k)$ approximates the inverse of the Hessian of $\varphi(\lambda)$ and we have an asymptotic Newton method as defined in [8].

Conclusions

The importance of an algorithm having the asymptotic Newton structure comes about in the following manner. First, at each iteration k the multipliers are updated in a way that is approximately steepest ascent on the dual function. The rate of convergence of steepest ascent is linear with the rate improvement a function of the ratio of the difference of the largest and smallest eigenvalues of the Hessian with the sum of these eigenvalues. With the algorithm presented here we have approximately steepest ascent within a vector space translated by the size of the multipliers. Within the transformed space the Hessian more closely approximates the identity matrix as r_k is increased bringing the largest and smallest eigenvalues closer together. This means that the factor that governs the rate of convergence is improving as r_k increases. As a consequence an asymptotic Newton algorithm has a superlinear convergence rate.

Whether this algorithm is superior to the algorithms that do not have the second order differentiability property is not clear. There is more computational effort at each iteration because nonbinding constraints still can affect the calculation of the Hessian for the unconstrained minimization long after they would have dropped out with the other algorithms. The choice of algorithm should be dependent on whether the discontinuity is a problem. A mixed strategy is probably best where this penalty function is used if problems occur with the discontinuity during an iteration.

1. Bertsekas, Kimitri, "On Penalty and Multiplier Methods for Constrained Minimization," Department of Electrical Engineering, working paper, University of Illinois at Urbana - Champaign, April 1974.

2. Bertsekas, Dimitri, "Combined Primal - Dual and Penalty Function Methods for Constrained Minimization," SIAM J. Control, Vol. 13, No. 3, 1975.

3. Evans, J. P. and F. J. Gould, "An Existence Theorem for Penalty Function Theory," SIAM J. Control, Vol. 12, No. 3, 1974.

4. Fiacco, A. V. and G. P. McCormick, Nonlinear Programming: Sequential Unconstrained Minimization Techniques, Wiley, New York, 1968.

5. Hestines, M. R., "Multiplier and Gradient Methods," J. Optimization Theory and Appl. 4 (1969), pp. 303-320.

6. Luenberger, D. G., Introduction to Linear and Nonlinear Programming, Addison-Wesley, Reading, Mass., 1973.

7. Mangasarian, O. L., Nonlinear Programming, McGraw-Hill, New York, 1969.

8. Mangasarian, O. L., "Unconstrained Methods in Optimization," Computer Sciences Technical Report #224, University of Wisconsin, 1974.

9. Murphy, F. H., "A Class of Exponential Penalty Functions," SIAM J. Control, Vol. 12, No. 4, 1974.

10. Murphy, F. H., "Topics in Nonlinear Programming, Penalty Function and Column Generation Algorithms," Ph.D. thesis, Yale University, 1971.

11. Murphy, F. H., "A Generalized Lagrange Multiplier Function Algorithm for Nonlinear Programming," Discussion paper 114, Center for Mathematical Studies in Economics and Management Science, October 1974.

12. Powell, M. J. D., "A Method for Nonlinear Constraints in Minimization Problems," in Optimization, R. Fletcher (ed.) Academic Press, New York, 1969.