Applied Mathematics NOTES de Mathématiques Appliquées

May 1980 Vol. 5, No.2 Mai 1980

CONTENTS CONTENU

Articles articles

Education Notes notes éducationelles

APPLIED MATHEMATICS NOTES

NOTES DE MATHÉMATIQUES APPLIQUÉES

Editor Rédacteur

W.R.S. Sutherland
Department of Mathematics
Dalhousie University
Halifax, Nova Scotia, B3H 4H8

The Applied Mathematics Notes does not publish technical research papers designed for specialized audiences. We do solicit carefully written manuscripts on the following topics:

Expository articles
Unusual mathematical models
Applications of mathematics
New curricula in applied mathematics
Education notes
Listings of preprints or technical reports

Les Notes de Mathématiques Appliquées ne publient pas les travaux techniques s'adressant aux spécialistes. La rédaction recherche des travaux soigneusement rédigés, tels

articles généraux
modèles mathématiques rares
applications nouvelle
programmes nouveaux en mathématiques appliquées
notes éducationelles
liste d'articles á paraître

Published by: Publiée par:

Canadian Mathematical Society
Société Mathématique du Canada
577 King Edward Avenue
Ottawa, Ontario, K1N 6N5
Canada

ISSN 0700-9224

Convex Programs with Equivalent Duals

Henry Wolkowicz
Department of Mathematics
The University of Alberta
Edmonton, Alberta, T6G 2G1

In this paper we study some recent duality results for the convex programming problem. We consider both the case of the ordinary convex program (P_o), which consists in minimizing a convex objective function subject to a finite number of convex inequality constraints, and the case of the abstract convex program (P), which has a cone constraint. Unlike usual duality results, e.g. [8], [9], the results presented here do not require that the constraints satisfy a regularity condition or constraint qualification. In addition, we point out the differences that arise between the ordinary convex program, the abstract convex program, and the presence or absence of a constraint qualification. A general characterization of optimality and a duality result, which cover all the mentioned cases, are given by (12) and the program (D) below. Several examples are included.

Duality results have proven extremely useful in mathematical programming, both computationally and theoretically. Many computational procedures begin by transforming the convex program into its dual program. In addition, the solution of the dual provides information for sensitivity analysis and has an economic interpretation as marginal values or shadow prices. The reader is referred to [2], [8], [9] for more details.

Let us first consider the ordinary convex program

$$(P_o) \qquad \mu = \underline{\inf}\ p(x)\ \underline{\text{subject to}}\ \ g^k(x) \le 0\ ,\ k = 1,\dots,m\ ,$$

Received Nov. 1979, revised Feb. 1980.

where $p: X \to R$ is a differentiable convex functional on the vector space X, while $g^k: X \to R$, $k = 1,\ldots,m$ are analytic convex functionals on X. If we let the vector function

$$g(x) = \left(g^k(x)\right)$$

and

$$S = R^m_+ \quad ,$$

be the nonnegative orthant in $Y = R^m$, then we can rewrite (P_o) in the form of the abstract convex program

$$(P) \qquad \mu = \underline{\inf}\ p(x) \ \ \underline{\text{subject to}}\ \ g(x) \in -S \ ,$$

where, by the convexity of the constraints g^k, it is easy to see that g is S-convex, i.e., for $x,y \in X$

$$(1) \qquad tg(x) + (1-t)g(y) - g(tx+(1-t)y) \in S \ , \qquad \text{for all} \ \ 0 \le t \le 1 \ .$$

Thus, the abstract convex program (P) generalizes the ordinary convex program (P_o). We now let S be any convex cone in R^m, i.e.

$$(2) \qquad S + S \subset S \ ; \ \lambda S \subset S \ , \qquad \text{for all} \quad \lambda \ge 0 \ ,$$

and we let g be any S-convex function which is also weakly analytic, i.e.

$$\phi g \text{ has a Taylor series, for all } \phi \text{ in } R^m \ ,$$

where ϕg is the dot product of ϕ and g in R^m. We let

$$S^+ = \{\phi \in R^m: \ \phi s \ge 0 \ , \ \text{for all} \ \ s \in S\}$$

be the nonnegative dual (apolar) cone of S.

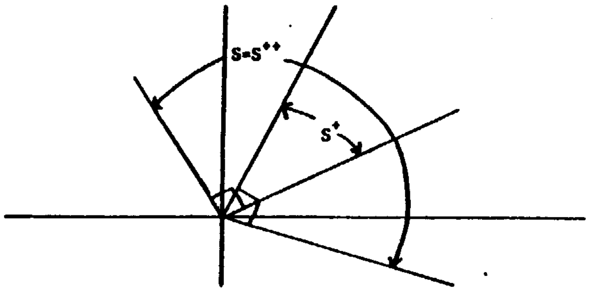


Figure 1. The apolar cone of the cone S

Note that if S is a closed convex cone, then g is S-convex if and only if the functions ϕg are convex for all $\phi \in S^+$. This follows from the fact that

$$S = S^{++} \ , \quad \text{for all closed convex cones} \ \ S \ .$$

This implies that if S is a polyhedral cone, i.e.

$$S = \bigcap_{i=1}^{t} \phi_i^+ \ , \quad \text{for some} \ \ \phi_i \ \ \text{in} \ \ R^m \ ,$$

then the abstract program (P) reduces to the ordinary convex program but with the t convex constraints $\phi_i g$, $i = 1,\dots,t$.

If some constraint qualification is satisfied for (P) , e.g. if Slater's condition holds for (P): "there is some $\hat{x}$ in X for which

$$(3)\qquad g(\hat{x}) \in -\text{int } S \text{ "} ,$$

where int denotes interior, then we get the following characterization of optimality and dual program of (P) [8], [9]

$$(4)\qquad \begin{cases} x^* \text{ (feasible) is optimal for (P)} \\ \qquad \text{if and only if} \\ \mu = \inf\{p(x) + \lambda g(x):\ x \in X\} ,\ \text{for some } \lambda \in S^+ \text{ with } \lambda g(x^*) = 0 , \end{cases}$$

and

$$(\bar{D})\qquad \mu = \sup_{\lambda\in S^+}\ \inf_{x\in X} p(x) + \lambda g(x) .$$

Thus the dual programs of (P) and (P_0) are essentially the same when Slater's condition is satisfied. Note that if S is polyhedral, then S^+ is also polyhedral (with $S^+ = R^m_+$ if $S = R^m_+$) .

Duality results such as the above can be used as a basis for solving (P). For example, if we define the dual functional

$$\phi(\lambda_0) = \inf_{x\in X} p(x) + \lambda_0 g(x) \quad ,$$

then (P) is equivalent to

$$\mu = \underline{\sup}\ \phi(\lambda_0)\quad \underline{\text{subject to}}\quad \lambda_0 \in S^+ .$$

If the gradient of ϕ is available, then $\phi(\lambda_0)$ can be calculated. The complexity of the problem now depends on how well we can describe S^+ . We will see below that in the case when no constraint qualification holds for (P) , then the definition of the dual functional ϕ includes the added restriction $x \in \Omega \subset X$, while the multipliers λ are (in certain cases) restricted to a larger cone $(S^f)^+ \supset S^+$.

Example 1. [9] (Optimal Control) Consider a dynamic system evolving in time and governed by the set of differential equations

$$(5)\qquad \dot{x}(t) = A(t)x(t) + b(t)u(t)$$

where $x(t)$ is an $m\times 1$ state vector , $\dot{x}(t)$ is the corresponding vector of derivatives, $A(t)$ is an $m\times m$ matrix, $b(t)$ is an $m\times 1$ distribution matrix , and $u(t)$ is a scalar control in $L_2[t_0,t_1]$, the space of square (Lebesgue) integrable functions on the interval $[t_0,t_1]$. Given the initial state $x(t_0)$ we seek the control u_0 minimizing

$$J(u) = \frac{1}{2}\int_{t_0}^{t_1} u^2(t)dt \quad ,$$

while satisfying the terminal inequalities

(6) $$x(t_1) \geq c \ ,$$

where c is a fixed $m \times 1$ vector and $t_1 \geq t_0$ is fixed. This problem might represent the selection of a thrust program for a rocket which has initial altitude and velocity $x(t_0)$ and which must exceed certain altitude and velocity limits by the given time t_1. The vector $x(t)$ then represents the altitude and velocity at time t while $u(t)$ is the acceleration force and $J(u)$ is the energy expended. We can write the solution to (5) in the form

$$x(t_1) = \Phi(t_1,t_0)x(t_0) + \int_{t_0}^{t_1} \Phi(t_1,t\ )b(t)u(t)dt \quad ,$$

where Φ is the fundamental solution matrix of the corresponding homogeneous equation. We assume $\Phi(t_1,t)$ and $b(t)$ to be continuous. The constraint (6) can be expressed as $g(u) \in -S$, where

(7) $$g(u) = c - x(t_1) = c - \Phi(t_1,t_t)x(t_0) - \int_{t_0}^{t_1} \Phi(t_1,t\ )b(t)u(t)dt \quad ,$$

and $S = R_+^m$. Thus, if we set $p(u) = J(u)$, we have a convex program (P) with $u \in X = L_2[t_0,t_1]$. Let us assume that Slater's condition holds, i.e. there exists a control $\hat{u}$ in X such that the components of the vector $g(\hat{u})$ in (7) are all strictly less that 0. (In example 3 below, Slater's condition fails and S is not polyhedral.) Then the duality result yields

$$\mu = \sup_{\lambda \in (R_+^m)^+} \ \inf_{u \in X} \ p(u) + \lambda g(u)$$

$$= \sup_{\lambda = (\lambda_i) \geq 0} \ \inf_{u \in X} \int_{t_0}^{t_1} \tfrac{1}{2}u^2(t) - \lambda\Phi(t_1,t)b(t)u(t)dt + \lambda(c - \Phi(t_1,t_0)x(t_0) \quad .$$

The inner unconstrained minimization problem can be solved for fixed λ_0 by differentiation, to yield

$$u_0(t) = \lambda_0\Phi(t_1,t)b(t).$$

Substituting for u, we see that the duality result has reduced the optimal control problem to the simple finite dimensional maximization problem

$$\mu = \sup_{\lambda \geq 0} \ \lambda Q \lambda' + \lambda d$$

where $'$ denotes transpose,

$$Q = \frac{-1}{2}\int_{t_0}^{t_1} \Phi(t_1,t)b(t)b'(t)\Phi'(t_1,t)dt$$

and

$$d = c - \Phi(t_1,t_0)x(t_0) \quad .$$

When no constraint qualification holds for the ordinary convex program (P_0), we have the following characterization of optimality [1], [3], [4] .

$$(8)\quad \begin{cases} x^* \text{ (feasible) is optimal for } (P_o) \\ \qquad \text{if and only if} \\ \nabla p(x^*) + \sum\limits_{k\in P(x^*)\setminus P^=} \lambda_k \nabla g^k(x^*) \in \left(\bigcap\limits_{k\in P^=} D^=_k\right)^+ \\ \qquad \text{for some } \lambda_k \geq 0 \;, \end{cases}$$

where

$$D^=_k = D^=_{g^k}(x) = \{d:\ \text{there exists } \bar{\alpha} > 0 \text{ with } g^k(x+\alpha d) = g^k(x) ,$$
$$\text{for all } 0 < \alpha \leq \bar{\alpha}\}$$

is the cone of directions of constancy of g^k ,

$$P = \{1,\dots,m\} \;,$$

$$P(x) = \{k\in P:\ g^k(x) = 0\}$$

is the set of binding (active) constraints at x and

$$P^= = \{k\in P:\ g^k(x) = 0 \text{ , for all feasible } x\}$$

is the set of equality constraints. Note that $D^=_h = D^=_h(x)$ is a subspace independent of x if h is an analytic convex function, e.g. [3]. The following equivalent characterizations of optimality were derived in [11], [12] :

$$\begin{cases} x^* \text{ (feasible) is optimal for } (P_o) \\ \qquad \text{if and only if} \\ \nabla p(x^*) + \nabla \lambda g(x^*) \in \left(\bigcap\limits_{k\in P^=} D^=_k\right)^+ \text{ , for some} \\ \lambda \in R^m_+ \text{ with } \lambda g(x^*) = 0 \;, \\ \qquad \text{if and only if} \\ \nabla p(x^*) + \nabla \lambda g(x^*) \in (D^=_h)^+ \text{ , for some} \\ \lambda \in R^m_+ \text{ , with } \lambda g(x^*) = 0 \end{cases}$$

where $h = \sum\limits_{k\in P^=} \alpha_k g^k$ and $\alpha_k \geq 0$ are any nonnegative scalars with $\alpha_k > 0$ if g^k is not affine. This yields the following characterization of optimality and dual program of (P_o)

$$(9)\quad \begin{cases} x^* \text{ (feasible) is optimal for } (P_o) \\ \qquad \text{if and only if} \\ \mu = \inf\{p(x) + \lambda g(x):\ x \in \Omega = \bar{x} + D^=_h\} \text{ , for some} \\ \lambda \in R^m_+ \text{ with } \lambda g(x^*) = 0 \;, \end{cases}$$

and (where $\bar{x}$ is any feasible point of (P_o))

$$(D_o)\qquad \mu = \sup_{\lambda\in R^m_+} \inf_{x\in\Omega} p(x) + \lambda g(x) \text{ , } \Omega = \bar{x} + D^=_h \;.$$

We will now see that the program (P) has an equivalent characterization of optimality as in (9) and an equivalent dual to (D_o) . First we need some preliminary notions. K is a face of S if

$$x, y \text{ in } S , x + y \text{ in } K \text{ implies } x, y \text{ in } K .$$

By S^f we denote the minimal face of (P) [6], [10] , i.e. S^f is the smallest face of S which contains $-g(F)$, where F denotes the feasible set of (P) . We say that the face K is exposed if

$$K = S \cap (\phi)^{\perp} ,$$

for some ϕ in S^+ , where $(\phi)^{\perp} = (\phi)^+ \cap (-\phi)^+$ is the hyperplane obtained by taking the orthogonal complement of ϕ . Thus if S^f is exposed, then

$$S^f = S \cap (\phi)^{\perp} , \tag{10}$$

for some $\phi \in S^+$. However, even when S^f is not exposed, then [7]

$$S^f = \bigcap_{k=1}^{t} (\phi^k)^{\perp} \cap S , \tag{11}$$

for some $\phi^k \in (S^f)^+$ where, with $\phi_o = 0$ and for $i = 1,\ldots,t$,

$$S^f \subset (S \cap \phi_o^{\perp} \cap \phi_1^{\perp} \cap \ldots \cap \phi_i^{\perp}) , \phi_i \in (S \cap \phi_o^{\perp} \cap \phi_1^{\perp} \ldots \cap \phi_{i-1}^{\perp})^+ .$$

This result uses the fact that S is finite dimensional.

Example 2. [6] Let S_1 denote the "ice-cream" cone in R^3

$$S_1 = \{x = (x_1,x_2,x_3): x_1 + x_2 \geq 0 , 2x_1x_2 - x_3^2 \geq 0)$$

and let S_2 denote the convex cone generated by S_1 and the point (1,0,1). Then the nontrivial faces of S_1 are exactly the boundary rays and all the faces are exposed. In fact, S_1 is the set of all vectors in R^3 which make an angle of 45° or less with the vector $I = (1,1,0)$, i.e.

$$S_1 = \left\{ x: \frac{\langle x, I\rangle}{\|x\| \, \|I\|} \geq \frac{1}{\sqrt{2}} \right\}$$

$$= \left\{ x: \frac{x_1 + x_2}{(x_1^2 + x_2^2 + x_3^2)^{1/2}} \geq 1 \right\} .$$

Now suppose that K is the boundary ray generated by the nontrivial boundary point x of S_1 . Then (see Figure 2.) the vector $\phi = (2,2,0) - \frac{2}{\|x\|}x$ is in S_1 and

$$K = S_1 \cap \phi^{\perp} .$$

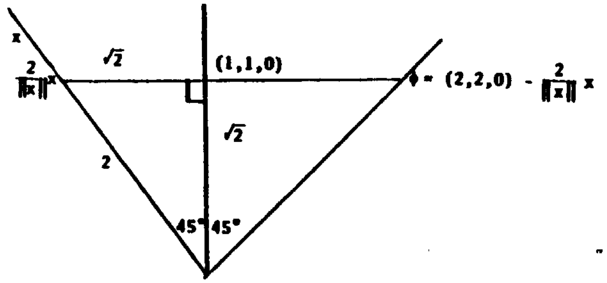


Figure 2

In S_2 , however, the ray through (1,0,0) , denoted K_1 , is extreme but not exposed and the smallest exposed face containing that ray is the convex cone generated by (1,0,0) and (1,0,1) (see Figure 3.)

$$K_2 = \text{cone } \{(1,0,0) \cup (1,0,1)\} \quad .$$

Note that $\phi_1 = (0,1,0)^t$ is in S_2^+ , for if $y \in S_2$, then, for some $\alpha, \beta \geq 0$ and $x = (x_i)$ in S_1 , we get

$$\phi_1 y = \phi_1(\alpha x + \beta(1,0,1)) = \alpha x_2 \geq 0 \ .$$

Furthermore

$$K_2 = S_2 \cap \phi_1^\perp \quad .$$

For if $y = \alpha x + \beta(1,0,1)$ for some $\alpha, \beta \geq 0$ and x in S_1 then $y = (y_i)$ is in $\phi_1^\perp \cap S_2$ if and only if

$$y_2 = \alpha x_2 = 0 \ ; \quad x_1 x_2 \geq x_3^2 \quad ; \quad x_1 + x_2 \geq 0$$

if and only if $y = \alpha x_1(1,0,0) + \beta(1,0,1)$, i.e. if and only if y is in K_2 . Now $\phi_2 = (0,0,1)$ is in K_2^+ and (as in (11))

$$K_1 = S_2 \cap \phi_1^\perp \cap \phi_2^\perp \quad .$$

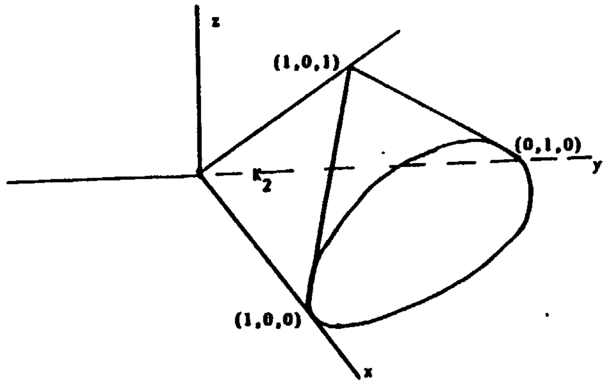


Figure 3

A characterization of optimality for (P) is [7]

$$(12) \quad \begin{cases} x^* \ \text{(feasible) is optimal} \\ \quad \text{if and only if} \\ \mu = \inf\{p(x) + \lambda g(x) : \ x \in \Omega = \overline{x} + \bigcap_{k=1}^{t} D_{h^k}^{=}\} \ , \\ \text{for some } \ \lambda \in (S^f)^+ \ \text{ with } \ \lambda g(x^*) = 0 \ . \end{cases}$$

Here $h^k = \phi^k g$, the ϕ^k satisfy (11) and $\bar{x}$ is feasible. Moreover, in the case that

$$(13) \qquad S^+ + (S^f)^\perp = (S^f)^+ \quad ,$$

then we can choose the multiplier λ in (12) to be in S^+ rather than $(S^f)^+$. Thus when (13) holds and S^f is exposed, we get the equivalent conditions to (9) .

The dual program of (P) is (where $\bar{x}$ is any feasible point of (P))

$$(D) \qquad \nu = \sup_{\lambda \in (S^f)^+} \inf_{x \in \Omega} p(x) + \lambda g(x) \ , \ \Omega = \bar{x} + \bigcap_{k=1}^{t} D^=_{h^k} \quad .$$

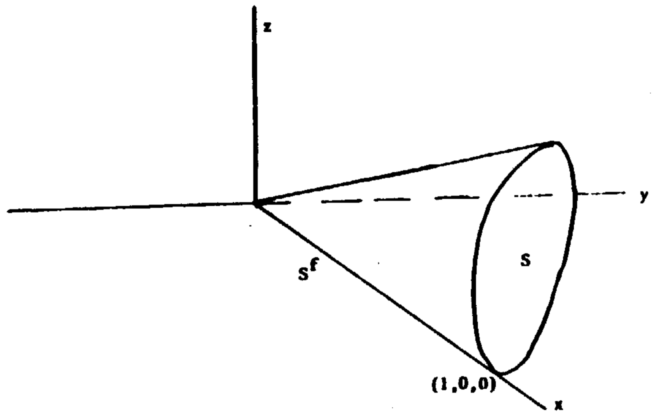
<u>Example 3.</u> Suppose that we consider the optimal control problem presented in Example 1 with $m = 3$, $t_0 = 0$, $t_1 = 1$, $x_0 = (1,1,1)$, $c = (4,1,1)$,

$$A(t) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad , \quad b(t) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad S = S_1 \ , \ \text{the "ice-cream" cone in Example 2.}$$

The constraint (6) is now

$$g(u) = c - x(1) \in -S$$

or equivalently, if $x(t) = (x_i(t))$,



<u>Figure 4</u>

$$x_1(1) \geq 4 \ ; \ x_2(1) \geq 1; \ 2x_1(1)x_2(1) \geq (x_3(1)-1)^2 \quad .$$

As in Example 1, x_1 might be the velocity, x_2 the altitude and x_3 might be the acceleration of the rocket. The fundamental solution matrix $\Phi(t_1,t) = e^{A(t_1,t)}$ and therefore, by direct calculation,

$$g(u) = (2 - \int_0^1 u(\tau)d\tau \ , \ 0 \ , \ 0 \ ) \ ;$$

$$S^f = S \cap \text{span } \{(1,0,0)\} \quad .$$

Moreover, it can be shown [6] that (13) holds and $S^+ = S$. Now the vector $\phi = (0,1,0)$ in S exposes the face S^f , i.e. $\phi^\perp \cap S = S^f$ and $D^=_{\phi g} = X$, since ϕg is identically zero. Therefore the characterization of optimality in (12) gives u^* (feasible) is optimal if and only if

$$0 = \nabla J(u^*) + \nabla\lambda g(u^*) = u^* - \lambda_1$$

with $\lambda = (\lambda_1) \in S$ and $\lambda g(u^*) = 0$,

i.e. if and only if u^* is the constant nonnegative function λ_1 and

$\lambda_1(2 - \int_0^1 u^*(\tau)d\tau) = 0$. This yields

$$u^* = u^*(t) = 2 .$$

Let us now summarize the results in this paper. We have compared the dual programs for the ordinary convex program (P_0) and the abstract convex program (P) . This was done in the presence and absence of constraint qualifications. The general form for the dual is given by the program (D) above. Let us point out the differences that arise in the various situations that we have considered.

(i) If we consider (P) as given by the ordinary convex program (P_0) , then $S = S^+$ is the nonnegative orthant in R^m (a polyhedral cone) and (13) always holds. Thus we can choose R^m_+ in (D) , i.e. the multipliers λ are restricted to be ≥ 0 . Moreover, every face of a polyhedral cone is exposed. This implies that $t = 1$. Finally, if some constraint qualification holds, then $\Omega = X$, i.e. the variable x is unrestricted.

(ii) If we consider (P) but with S polyhedral, then we have the same situation as in (i) above, but with the exception that the multipliers λ are restricted to the dual cone (still polyhedral) S^+ .

(iii) If relation (13) fails, then the sup is taken with λ restricted to the larger cone $(S^f)^+$ rather than S^+ .

(iv) If the minimal cone S^f is exposed, then we have $t = 1$. (This always occurs if S is polyhedral or for example if $S = S_1$ in Example 2).

(v) If some constraint qualification holds, then we can restrict the multipliers λ to S^+ and $\Omega = X$. Thus (D) is equivalent to $(\bar{D})$ in this case.

In conclusion, let us note that many of our assumptions can be relaxed. First, we do not need ϕg analytic for all $\phi \in R^m$, but rather that $\phi^k g$ be analytic for $k = 1,\ldots,t$, where the ϕ^k satisfy (11). This guarantees that the functions $\phi^k g$ are faithfully convex , i.e. they are affine on a line segment only if they are affine on the whole line containing that segment. This then implies that the cone of directions of constancy $D^{=}_{\phi^k g}$ is a subspace independent of x in X . If $X = R^n$, then we can calculate the cones $D^{=}_{\phi^k g}$ [13] . The assumption of differentiability of the objective function p and the constraint g can also be relaxed and, in particular, can be replaced by continuity when the $\phi^k g$, $k = 1,\ldots,t$ are faithfully convex. The optimality conditions are, in this case, given using subdifferentials.

References

[1] R.A. Abrams and L. Kerzner, "A simplified test for optimality", Journal of Optimization Theory and Applications 25 (1978), 161-170.

[2] M.L. Balinsky and W.J. Baumol, "The Dual in Nonlinear Programming and its Economic Interpretation", The Review of Economic Studies 35 (1968), 237-256.

[3] A. Ben-Israel, A. Ben-Tal and S. Zlobec, "Optimality conditions in convex programming", The IX Internation Symposium on Mathematical Programming, Budapest (1976).

[4] A. Ben-Israel, A. Ben-Tal and S. Zlobec, Optimality in Nonlinear Programming: a feasible directions approach, J. Wiley, London New York-Sydney (forthcoming).

[5] J. Borwein and H. Wolkowicz, "Characterization of optimality without constraint qualification for the abstract convex program", Research Report No. 14, Dalhousie University (1979).

[6] J. Borwein and H. Wolkowicz, "Characterizations of optimality for the abstract convex program", Research Report No. 19, Dalhousie University (1979).

[7] J. Borwein and H. Wolkowicz, "Regularizing the abstract convex program", The University of Alberta (1979).

[8] A.M. Geoffrion, "Duality in nonlinear programming: a simplified applications--oriented development", SIAM Review 13 (1971) 1-37.

[9] D.G. Luenberger, "Optimization by vector space methods, J. Wiley & Sons (1969).

[10] H. Massam, "Optimality conditions for a cone-convex programming problem", Journal of Australian Mathematical Society (Series A) 27 (1979).

[11] H. Wolkowicz, "Geometry of optimality conditions and constraint qualifications: the convex case", Mathematical Programming (in press).

[12] H. Wolkowicz, "A strengthened test for optimality", The University of Alberta (1979).

[13] H. Wolkowicz, "Calculating the cone of directions of constancy", Journal of Optimization Theory and Applications 25 (1978) 451-457.

THE MATHEMATICS OF DEMOCRACY

Wayne Patterson
Professeur agrégé
Département de physique-mathématiques
Université de Moncton
Moncton, N.-B., E1A 3E9

Résumé: Les Mathématiques de la Démocratie

Considérons un ensemble fini de cardinalité P , et une partition de P dans M sous-ensembles. Choisissons de chaque sous-ensemble, de cardinalité P_i , N_i éléments $(i = 1,\ldots,M)$ tels que la fonction

$$F = \sum_{i=1}^{M} D(\frac{N}{P}, \frac{N_i}{P_i})$$

est minimisée sujet à la contrainte $\Sigma N_i \leq N$. (D = distance.)

Ce problème peut être interprete dans une façon qui rend sa solution bien plus intéressante que son contenu intrinsèque. Dans son application à la vie, elle peut être exprimée:

"Comment peut un corps représentatif (comme le Parlement) diviser un nombre de membres fixe entre un nombre de juridictions, de façon qu'un seul représentant ne peut pas représenter les parties de deux juridictions, telles que la population moyenne de chaque région (comme une circonscription) est si proche à la moyenne générale comme possible?"

Ce problème arrive, au Canada et aux Etats-Unis, avec chaque recensement décennal, lors de la redistribution des bornes électorales. Jusqu'à présent, aucun essai à formuler le problème mathématique associé s'est produit au Canada. Dans les Etats-Unis, une fonction appelée la "Méthode des Proportions Egales" est employée pour calculer la redistribution.

Notre présentation démonstrera que la fonction utilisée aux Etats-Unis ne minimise pas les inégualités entre les régions, et démonstrera aussi un algorithme pour déterminer les valeurs minimales pour F .

INTRODUCTION

More and more today, we see problems that humankind encounters, in whatever sphere of activity, being subjected to various forms of systematic, i.e. mathematical, analysis. There have been mathematics papers written on subjects as diverse as the Fundy tides and the scheduling of pro football games.

Many of these problem areas invade what is called "public sector" activity; they aid in making decisions for the body politic. However, it is interesting that a mathematical problem which is at the heart of the body politic has rarely been discussed in mathematical terms, and, in the forum of politics, not since 1941; and, the mathematical solution described then is sadly lacking in the context of our present-day understanding of the power of mathematics.

THE PROBLEM (STATED NON-MATHEMATICALLY)

Many countries are governed by a legislative body that is directly responsible to the people. That is, in such countries, when an elector goes to the ballot box, he or she casts a vote for a certain person, who, if elected, will sit in the legislative body as a representative of that elector and all the other electors of his or her district.

This system, called representative democracy, differs, say, from a system of proportional representation, wherein, if

36% of the people cast votes for party X or leader Y, then that party or leader may select 36% of the representatives in the legislative body.

Our problem arises in representative democratic systems, where there is more than one level of government, and the method of election to the higher level respects the boundaries of the lower.

For example, in Canada we have the federal government and ten provinces, and by the BNA Act and tradition no federal member of parliament may represent parts of two provinces — say, for example with a riding consisting of Amherst, N. S., and Sackville, N. B.

The problem, which exists as well in the United States, Australia, France, and West Germany, among others [4], is that there is an (integral) number of seats, N, in the national legislature (N = 282 in Canada, N = 435 in the United States) and an (integral) number from each province; but the proportion of the total population in each province is unlikely to give a partition of N into integral parts proportionally.

Furthermore, it has been established, judicially in the United States and by tradition in Canada, that one person's vote, from wherever in the country, should count equally with that of anyone else — the tradition of one man, one vote; which, thankfully today, can be restated as one person, one vote.

Thus by this principle, it is necessary to create legislative districts which are as equal in size as possible while respecting the boundaries of the lower jurisdictions of government (such as provinces or states).

THE PROBLEM (STATED MATHEMATICALLY)

Consider a finite set of P elements, and a partition of P into M subsets. Select from each subset of cardinality P_i, N_i elements $(i = 1, \ldots, M)$ such that the function

$$F = \sum_{i=1}^{M} D\left(\frac{N}{P}, \frac{N_i}{P_i}\right)$$

is minimized subject to the constraint $\sum_{i=1}^{N} N_i \leq N$.

Here P represents the total population of a country; P_i, the population of the i^{th} province; N, the total number of representatives in the legislative body; and N_i, the number of representatives from province i. D is a distance function measuring the distance of the average representation (per person) in a riding in province i from the national average representation (per person). F is the sum of these distances for all states or provinces.

Otherwise stated, we have the following problem. How can a representative body divide a fixed membership among a number of jurisdictions, so that the average population in any representative's district is as close to the average as possible, that is, so that the function F is minimized?

HISTORY

The author's own curiosity was piqued earlier last year when he discovered an article describing projections for the 1980 U. S. census, and the effects that this census would be likely to have on the state-by-state representation in Congress.

The Congress [1] decided in the 1920's to fix the representation of the House of Representatives at 435. (The United States Senate is fixed at 100, but does not follow the one-man-one-vote principle, but rather, the one-state-two-vote principle.) A number of methods had been used to arrive at the representation, but in 1941 a method called the Method of Major Fractions was in use. It had been refined by a Professor Willcox of the Cornell University Mathematics Department.

But in contention was a method called the Method of Equal Proportions, proposed in the 1920's by a Professor Huntington of the Harvard Mathematics Department, and also recommended by the U. S. National Science Foundation [3].

In 1941, the (friendly ?) academic debate between Harvard and Cornell broke out into full-scale war in the halls of Congress. It appeared that, using the 1940 Census and the Harvard method, that one district more would be given to Arkansas and taken away from Michigan, than would have been the case with the Cornell method [6].

All of the other 46 (at that time) states would have had the same representation using either method.

There followed a fascinating debate stretching over several months, and mostly involving — as one might imagine — the Congressmen from Arkansas and Michigan.

One is treated to the remarkable spectacle of Arkansas Congressmen arguing about the mathematical superiority of the Harvard method, and similarly for Michigan and Cornell. Finally, the veneer is stripped away and I quote this fascinating passage [2]:

> Senator Brown (of Michigan): Mr. President, I cannot refrain mentioning now a subject which I know was a potent force in the House of Representatives and is a potent force in the Senate of the United States. That is the politics of this situation. In the House of Representatives one Republican voted for the apportionment bill. In the House, outside of the State of Michigan, ..., only three Democrats voted against the bill.
>
> It was a strict party vote. What was the reason for that? The reason was that Arkansas is considered to be a sure Democratic State, ... and because Michigan is considered to be a doubtful state. I wish to inquire into that situation and analyze it briefly from a party standpoint, and I want to appeal to the reasoning of my Democratic colleagues in the Senate upon that subject.

Eventually, because of the Rooseveltian Democratic majority, Equal Proportions was adopted and is in effect today.

THE METHOD OF MAJOR FRACTIONS

The Method of Major Fractions, a refinement of the "intuitive" solution, is the following: Assign to each state one representative. This is constitutionally guaranteed. For the remaining 385, first construct a sequence for each state, S_i, with population P_i:

$$S_i = \{P_i,\ \frac{1}{3}P_i,\ \frac{1}{5}P_i,\ \dots,\ 1/(2N-1)\,P_i,\ \dots\}$$

Let $S = \bigcup_{i=1}^{50} S_i$. Choose the 385 largest elements of S. If S_i has N_i of these largest elements, then, in all, state S_i will be allocated $N_i + 1$ representatives.

THE METHOD OF EQUAL PROPORTIONS

The Method of Equal Proportions is similar, except that the sequence $S_i' = \{P_i,\ 1/\sqrt{2}\,P_i,\ 1/\sqrt{6}\,P_i,\ 1/\sqrt{3 \cdot 4}\,P_i,\ \dots,\ 1/\sqrt{N(N-1)}\,P_i,\ \dots\}$ is used.

Throughout the literature of the day one reads [6] that Equal Proportions was fairer because if two states exchanged a representative after a census, the relative proportions of the state average populations per district would be closer to 1.

Theorem 1. Harvard was partly right.

Proof: If S_i gains a representative from S_j under Equal Proportions, then at least one of the elements of S_i' will now be in the "Top 385" at the expense of one of the elements of S_j'. Suppose that S_i goes from N_i to $N_i + 1$ and S_j from N_j to $N_j - 1$.

I will demonstrate only the case $P_i > P_j$ (hence $N_i \geq N_j$), and $\frac{P_j \cdot N_i}{P_i \cdot N_j} < 1$.

It is necessary to show that $\frac{P'_j \cdot (N_i + 1)}{P'_i \cdot (N_j - 1)}$ is closer to 1 than is $\frac{P_j \cdot N_i}{P_i \cdot N_j}$.

Before the census is taken, the "cut-offs" are N_i and N_j so that,

$$P_i \sqrt{\frac{1}{N_i(N_i - 1)}} > P_j \sqrt{\frac{1}{N_j(N_j - 1)}} > P_i \sqrt{\frac{1}{N_i(N_i + 1)}}$$

which leads to

$$\sqrt{\frac{N_i(N_j - 1)}{N_j(N_i - 1)}} > \frac{P_j \cdot N_i}{P_i \cdot N_j} > \sqrt{\frac{N_i(N_j - 1)}{N_j(N_i + 1)}}$$

After the census, we have

$$P'_i \sqrt{\frac{1}{N_i(N_i + 1)}} > P'_j \sqrt{\frac{1}{N_j(N_j - 1)}} > P'_i \sqrt{\frac{1}{(N_i + 1)(N_i + 2)}}$$

or

$$\sqrt{\frac{N_j(N_i + 1)}{N_i(N_j - 1)}} > \frac{P'_j \cdot (N_i + 1)}{P'_i \cdot (N_j - 1)} > \sqrt{\frac{N_j(N_i + 1)}{(N_i + 2)(N_j - 1)}}$$

Since $N_i \geq N_j$, we have

$$\sqrt{\frac{N_j(N_i + 1)}{(N_j - 1)(N_i + 2)}} > \sqrt{\frac{N_i(N_j - 1)}{N_j(N_i - 1)}}$$

as we wished to prove.

Showing that $\frac{P'_j \cdot (N_i + 1)}{P'_i \cdot (N_j - 1)}$ is bounded by 1 follows similarly.

<u>Theorem 2.</u> Harvard was mostly wrong.

Proof: The Method of Equal Proportions may, in fact, worsen the proportions between all pairs of states. For example, as long as $P/N >> M$ (e.g. in the U. S. $P/N \approx 500{,}000$; $M = 50$); then, let the states be ordered by population averages P_i/N_i so that $P_1/N_1 \geq P_2/N_2 \geq \ldots \geq P_{50}/N_{50}$.

Then a new census P'_i such that

$$P'_i = P_i + M - i$$

will ensure that the ratios $\frac{P'_i \cdot N_j}{P'_j \cdot N_i}$ will be further from 1 than their predecessors.

THE CANADIAN CONTEXT

Here in Canada we have not even achieved the sophistication of the Americans. In 1974, the Parliamentary Standing Committee on Privileges and Elections was presented with five calculation methods for the redistribution of Parliament. They were named, rather exotically [8] the Compensation Method, Quebec Plus Four, Qualified Parity, Amplified, and Amalgam. None of these methods, however, were more than variations on the "intuitive method", except with different approaches for keeping the minimum legislative guarantees in Canada of no province having fewer Commons seats than Senate seats, and no province having $N_i > N_j$ if $P_i \leq P_j$.

THE SOLUTION

Since there are only finitely many choices for $\{N_i\}$, in fact with an upper bound of N^m, then there must exist a set of values $\{N_i^*\}$ which minimize F.

The problem, as stated, is a straightforward problem in dynamic linear programming, with F to be optimized subject only to the constraint that $\sum_{i=1}^{M} N_i = N$. The solution algorithm is described, for example, in [5]. The following tables list a number of cases that have been tested using the dynamic programming algorithm DYNPR, written in the APL language, at the University of California (Riverside).

A word is necessary about the distance function, D. In applying optimization techniques, some notion of "distance" from the average must be chosen. Some obvious candidates are:

$$D(\frac{N_i}{P_i}, \frac{N}{P}) = \left|\frac{N_i}{P_i} - \frac{N}{P}\right|$$

$$D(\frac{N_i}{P_i}, \frac{N}{P}) = (\frac{N_i}{P_i} - \frac{N}{P})^2$$

$$D(\frac{N_i}{P_i}, \frac{N}{P}) = (\frac{N_i}{P_i} - \frac{N}{P})^K$$

One might also sum over the number of people in the country rather than the number of jurisdictions; thus, for example, we could have

$$F = \sum_{i=1}^{M} (\frac{N_i}{P_i} - \frac{N}{P})^2$$

or

$$F = \sum_{1}^{N} (\frac{N_i}{P_i} - \frac{N}{P})^2$$

$$= \sum_{1}^{M} P_i \cdot (\frac{N_i}{P_i} - \frac{N}{P})^2 .$$

Herein are listed the dynamic programming solutions to a number of redistribution problems for the Canadian House of Commons:

Case # 1	1951 census - 265 members of parliament
Case # 2	1961 census - 264 members of parliament
Case # 3	1971 census - 282 members of parliament
Case # 4	1976 census - 282 members of parliament

Each case is divided into four subcases, the results with respect to each of the following distance functions:

(a) $F_a = \sum_{i=1}^{M} (\frac{N_i}{P_i} - \frac{N}{P})^2$

(b) $F_b = \sum_{i=1}^{M} P_i(\frac{N_i}{P_i} - \frac{N}{P})^2$

(c) $F_c = \sum_{i=1}^{M} \left|\frac{N_i}{P_i} - \frac{N}{P}\right|$

(d) $F_d = \sum_{i=1}^{M} P_i\left|\frac{N_i}{P_i} - \frac{N}{P}\right|$

CASE # 1 CANADA - 1951 CENSUS - 265 MP's

PROVINCE	POPULATION	ACTUAL SEATS	F(a)	F(b)	F(c)
Newfoundland	361,416	7	7	7	7
Prince Edward I.	98,429	4	4	4	4
Nova Scotia	642,584	12	12	12	12
New Brunswick	515,697	10	10	10	10
Quebec	4,055,681	75	75	76	76
Ontario	4,597,542	86	84	86	78
Manitoba	776,541	14	15	14	15
Saskatchewan	831,728	17	16	15	16
Alberta	939,501	17	18	17	18
British Columbia	1,165,210	23	22	23	23
Yukon	9,096	1	1	1	1
NWT	16,004	1	1	1	1
Total	14,009,429	265	265	265	260
$F_{(actual)}$			10.65×10^{-9}	.00015	.00016
F			10.65×10^{-9}	.00015	.00016

CASE # 2 CANADA - 1961 CENSUS - 264 MP's

PROVINCE	POPULATION	ACTUAL SEATS	F(a)	F(b)	F(c)
Newfoundland	457,853	7	7	7	7
Prince Edward I.	104,629	4	4	4	4
Nova Scotia	737,007	11	11	10	11
New Brunswick	597,936	10	10	10	10
Quebec	5,259,211	74	74	75	76
Ontario	6,236,092	88	88	88	86
Manitoba	921,686	13	13	13	13
Saskatchewan	925,181	13	13	13	13
Alberta	1,331,944	19	19	19	19
British Columbia	1,629,082	23	23	23	23
Yukon	14,628	1	1	1	1
NWT	22,998	1	1	1	1
Total	18,238,247	264	264	264	264
$F_{(actual)}$			4.3148×10^{-9}	.00012643	.000112292
F			4.3148×10^{-9}	.00012642	.000112232

CASE # 3 CANADA - 1971 CENSUS - 282 MP's

PROVINCE	POPULATION	ACTUAL SEATS	F(a)	F(b)	F(c)	F(d)
Newfoundland	522,104	7	7	7	7	6
Prince Edward I.	111,641	4	4	4	4	4
Nova Scotia	788,960	11	10	10	10	10
New Brunswick	634,557	10	10	10	10	10
Quebec	6,027,764	75	77	77	78	78
Ontario	7,703,106	95	98	97	95	100
Manitoba	988,247	14	13	12	13	12
Saskatchewan	926,242	14	12	12	12	12
Alberta	1,627,874	21	21	21	21	21
British Columbia	2,184,621	28	28	28	28	27
Yukon	18,388	1	1	1	1	1
NWT	34,807	2	1	1	1	1
Total	21,568,311	282	282	280	280	282
$F_{(actual)}$			4.2095×10^{-9}	.000175	.000117	20.7498
F			2.47820×10^{-9}	.000106	.000008	11.0963

CASE # 4 CANADA - 1976 CENSUS - 282 MP's

PROVINCE	POPULATION	ACTUAL SEATS	F(a)	F(b)	F(c)	F(d)
Newfoundland	562,500	7	7	7	7	6
Prince Edward I.	120,300	4	4	4	4	4
Nova Scotia	835,400	11	10	10	10	10
New Brunswick	686,400	10	10	10	10	10
Quebec	6,283,100	75	74	75	76	76
Ontario	8,373,500	95	99	100	85	101
Manitoba	1,031,300	14	12	12	12	12
Saskatchewan	936,500	14	11	11	11	11
Alberta	1,899,700	21	23	22	23	23
British Columbia	2,497,600	28	30	29	30	27
Yukon	21,500	1	1	1	1	1
NWT	43,300	2	1	1	1	1
Total	23,291,100	282	282	282	270	282
$F_{(actual)}$			2.8×10^{-9}	1.0×10^{-6}	1.0×10^{-6}	23.25
F			1.75×10^{-9}	0.90×10^{-6}	0.75×10^{-6}	11.48

REMARKS ON THE CASE STUDIES

In each case analysed, the "constitutional minima" were built into the program. In other words, the variables used were

$$N'_{\text{NEWFOUNDLAND}} = N_{\text{NEWFOUNDLAND}} - 6$$

$$N'_{\text{PRINCE EDWARD ISLAND}} = N_{\text{PRINCE EDWARD ISLAND}} - 4$$

$$N'_{\text{NOVA SCOTIA}} = N_{\text{NOVA SCOTIA}} - 10$$

$$N'_{\text{NEW BRUNSWICK}} = N_{\text{NEW BRUNSWICK}} - 10$$

$$N'_{\text{QUEBEC}} = N_{\text{QUEBEC}} - 24$$

$$N'_{\text{ONTARIO}} = N_{\text{ONTARIO}} - 24$$

$$N'_{\text{MANITOBA}} = N_{\text{MANITOBA}} - 6$$

$$N'_{\text{SASKATCHEWAN}} = N_{\text{SASKATCHEWAN}} - 6$$

$$N'_{\text{ALBERTA}} = N_{\text{ALBERTA}} - 6$$

$$N'_{\text{BRITISH COLUMBIA}} = N_{\text{BRITISH COLUMBIA}} - 6$$

$$N'_{\text{YUKON}} = N_{\text{YUKON}} - 1$$

$$N'_{\text{NORTHWEST TERRITORIES}} = N_{\text{NORTHWEST TERRITORIES}} - 1$$

where the constants represent the number of Senators per jurisdiction, and hence the minimum number of seats guaranteed to the jurisdiction.

Consequently, the constraint is transformed from

$$\Sigma N_i = M$$

to

$$\Sigma N_i' = M - S$$

where S represents the size of the Senate.

Define the variable Δ_Y to be the average, using distance functions $F_{(a)}$, $F_{(b)}$, $F_{(c)}$, and $F_{(d)}$ of the differences between the optimal solution and the actual solution.

For example, for 1951 we have

$$\Delta_{1951,}\ F_{(a)} = 6$$

$$\Delta_{1951,}\ F_{(b)} = 3$$

$$\Delta_{1951,}\ F_{(c)} = 12$$

$$\therefore\ \Delta_{1951} = 7\ .$$

In all, we have:

$$\Delta_{1951} = 7$$

$$\Delta_{1961} = 2$$

$$\Delta_{1971} = 11$$

$$\Delta_{1976} = 17.5$$

One might facetiously refer to Δ as a "gerrymandering coefficient". It provides a crude measure of the difference between the actual redistributions and the various optimal solutions.

On the surface, it certainly appears that efforts at redistribution in recent years have fallen further and further from the "one man, one vote" principle, since both the 1951 and 1961 redistributions were very close to the optimal solutions.

One would expect that the demographic reason for this comparatively recent departure from optimality is due to a more rapidly fluctuating distribution in Canadian population, and an inability of the present methods to reflect this fluctuation.

RECENT MATHEMATICAL DEVELOPMENTS

In the past few years, there has been mathematical development in the study of redistribution methods (although none of this development has been translated into public policy). Particularly, the work of Still [9] and Balinski and Young [10] should be noted in this regard.

However, this work is predicated upon the development of methods that introduce other constraints not considered in this paper.

For example, Still, Balinski and Young use the concepts of "quota" and "house monotonicity" in defining redistribution functions (called apportionment methods) as follows: A

redistribution $\{n_i\}$ satisfies "quota" if

$$\lfloor (\frac{N P_i}{P}) \leq n_i \leq \lceil (\frac{N P_i}{P})$$

where $\lfloor$ represents the greatest integer less than, and $\lceil$ the smallest integer greater than.

A redistribution is "house monotone" if an increase in N (to N') cannot result in $N_i' < N_i$ for the new redistribution $\{N_i'\}$.

These constraints are not considered in this paper as neither is consistent with the strictest possible application of the one-man, one-vote principle. A plan for further investigation includes the incorporation of the Still-Balinski-Young methods along with the optimization approach taken here.

REFERENCES

[1] "1980 Reapportionment May Take Two Seats from Pennsylvania", Roll Call, July 12, 1979.

[2] United States Congressional Record, p. 8078-79, October 21, 1941.

[3] Bibliography of Methods of Apportionment in Congress, American Mathematical Monthly 42, February 1942.

[4] Studies in Federalism, R. V. Bowie and C. J. Friedrich, Little Brown 1954.

[5] Mathematical Programming for Economics and Business, R. C. Pfaffenberger and D. A. Walker, Iowa State U. Press, 1976.

[6] The 1941 Apportionment Bill, Science News 93, No. 2411, p. 6, 1941.

[7] Preliminary Report, N. S. Select Committee on Electoral Boundaries, 1977.

[8] House of Commons, Standing Committee on Privileges and Elections, Minutes of Proceedings, April 9, 1974.

[9] "A Class of New Methods for Congressional Apportionment", Still, J. W., SIAM Review, October 1979, pp. 401-418.

[10] "A New Method for Congressional Apportionment", Balinski, M. L. and Young, H. P., Proc. Nat. Acad. Sci., U.S.A., 71 (1974) pp. 4602-4606.

ACKNOWLEDGEMENTS

The author is indebted to Professor E. Tory for bringing papers [9] and [10] to his attention.

An Erratum. (Volume 5, Number 1, Feb. 1980) page 15, second line from bottom: "then amounts to applying the usual non-slip condition ..."
should read
"then amounts to applying the usual no normal velocity condition ... ".

EDUCATION NOTES

A Regular Column on Simple but Interesting Problems in Differential Equations

Edited by

Frederic Y.M. Wan, Director,
Institute of Applied Mathematics & Statistics
University of British Columbia
Vancouver, B.C., V6T 1W5.

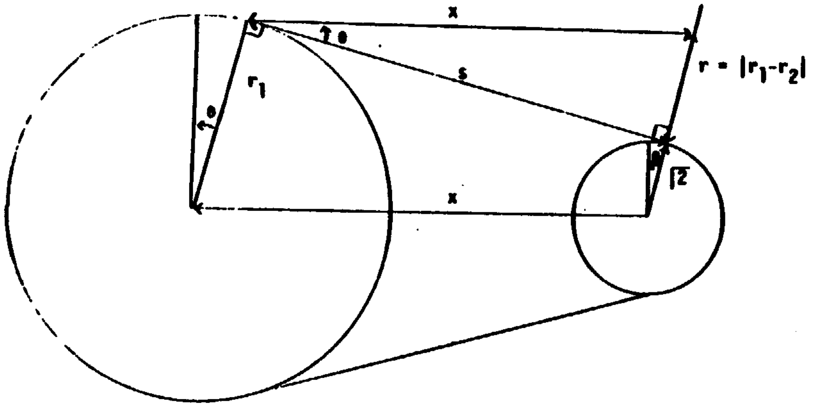
Note: Submissions to this column may be sent to Fred Wan or to the Editor of Applied Mathematics Notes. Material need not be restricted to differential equations.

OF BICYCLES AND NEWTON

Ronald W. Gatterdam
Mathematics and Computer Science
University of Wisconsin-Parkside
Kenosha, Wisconsin, 53141

Consider two sprocket wheels with a chain running between them as depicted in the sketch. A designer of such a system has the problem of adjusting the center-to-center distance, x , so that the length of the sprocket chain, ℓ , is equal to an integer number of chain links of length D . In the case of a bicycle, mechanical adjustment of x can be provided but if the sprocket wheels are attached to machinery it is often impractical to provide for much adjustment. Thus, the distance x must be computed with some accuracy. We show how this computation can be made by a simple Newton-like method.

Received (via J.W. Macki, Univ. of Alberta) Feb., 1980;
revised Apr. 1980.



Figure

First, to compute the length of the chain observe that the chain is tangent to the wheels at the point of contact. Also, note that by translating the center-to-center line segment along the parallel radii to the contact points, a right triangle with hypotenuse x and legs $r = |r_1 - r_2|$ and s, the tangential chain length, is obtained (the absolute value is chosen so it makes no difference which of the radii is larger). The angle, θ, between the hypotenuse and the leg of length s is also the angle between the radii tangential to the chain and those normal to the center line. The relations are:

$$r = |r_1 - r_2|$$
$$s = \sqrt{x^2 - r^2}$$
$$\theta = \tan^{-1}(r/s)$$
$$\ell = 2s + (r_1 + r_2)\pi + 2r\theta \ .$$

Here is where Newton plays a role. Given a nominal center-to-center distance x, the approximate number of links is given by $n = \text{int}(\ell/D + .5)$ (where int is the greatest integer function and the $+ .5$ is used to round to the nearest integer). The length x is to be adjusted so that $\Delta\ell = nD - \ell$ is small. Now

$$\frac{\Delta\ell}{\Delta x} \simeq \frac{d\ell}{dx}$$

so

$$\Delta x \simeq \Delta\ell/\frac{d\ell}{dx} \ .$$

Note that the sign of $\Delta\ell$ is chosen to agree with the direction in which change is required, i.e., ℓ is to be replaced by $\ell + \Delta\ell$. A few applications of the chain rule and some elementary algebra show

$$ds = (x/s)dx$$
$$d\theta = (-r/(xs))dx$$
$$d\ell = (2s/x)dx$$
$$\text{so} \quad \Delta x \simeq x\Delta\ell/(2s) \ .$$

The method is now easily described. Starting with an initial value of x, compute $s,\theta,\ell,n,\Delta\ell,\Delta x$ as indicated; replace x by $x + \Delta x$; repeat the process until $\Delta\ell$ is less than the prescribed accuracy.

It can be shown that if the initial value of s is not close to 0, the initial value of x is greater than r_1+r_2, and D is smaller than the minimum of r_1 and r_2, then the algorithm converges to the "nearest possible" solution. "Nearest possible" means that for ℓ_0 the initial chain length and ℓ_1 the chain length to which the algorithm converges, $|\ell_1 - \ell_0| < D/2$. (In fact the algorithm converges provided only that $s > 0$ and $\min(r_1,r_2) > 0$, but may converge to unusual or negative values for x.) Sufficient conditions for convergence to the nearest solution are that for x_0 the initial value of x, θ_0 the initial value of θ, and $\rho = \min(r_1,r_2)$, $x_0 > r_1+r_2$ and $D \leq 4\rho\cos^2\theta_0$. The interested reader may verify the above by considering the first two terms of the Taylor series for ℓ as a function of x with remainder.

Observe that the conditions expressed above are reasonable in the physical situation. In particular, the condition on D and θ_0 can be viewed as expressing the requirement that the chain make reasonable contact with the smaller sprocket. In practice the method converges rapidly to the nearest solution.

Some Sample Computations:

r_1	r_2	D	x	ℓ	n	Δℓ	Δx
10	2	.5	12.000 12.159 12.158	67.263 67.501 67.500	135 135 135	.237 -.001 0	.159 -.0008
24	3	.375	30.00 29.92	160.238 160.125	427 427	-.11 0	-.08
5	1	.375	25.000 24.941	69.49 69.375	185 185	-.1159 0	-.0587
10	1	.375	30.00 29.92	97.278 97.125	259 259	-.1533 0	-.0804

SUBSCRIPTIONS

The subscription rate for the Applied Mathematics Notes is $6.00 (Canadian) per annum. Subscriptions should be sent to the office of the Canadian Mathematical Society (see inside front cover).

ABONNEMENT

Le prix de l'abonnement (annuel) aux Notes de Mathématiques Appliquées est de $6.00 (Canadien). Faire parvenir paiement au bureau de la Société Mathématique du Canada (voir le revers de la page couverture).

INSTRUCTIONS TO AUTHORS

Since the Applied Mathematics Notes are printed by a photo-offset process, authors are requested to submit the original typescript (plus one copy) on 8 1/2 by 11 or A4 paper.

Printed at: Imprimé à:

Dalhousie University, Halifax, N.S. B3H 4H8