Mathematical
Inequalities
& Applications
Volume 7, Number 3 (2004), 309–313

A VARIATIONAL PROOF OF BIRKHOFF'S THEOREM ON DOUBLY STOCHASTIC MATRICES

QIJI J. ZHU

(*communicated by J. Borwein*)

Abstract. This note provides a short variational proof of the Birkhoff's theorem asserting that the extreme points of the convex set of doubly stochastic matrices are the permutation matrices.

1. Introduction

An N by N square matrix $A = (a_{nm})$ is *doubly stochastic* provided that the entries of A are all nonnegative, $\sum_{n=1}^{N} a_{nm} = 1$ for $m = 1, \ldots, N$ and $\sum_{m=1}^{N} a_{nm} = 1$ for $n = 1, \ldots, N$. Doubly stochastic matrices naturally arise in analyzing stochastic processes. They are also closely related to the concept of majorization which has important applications in physics and economics. Ando's survey paper [1], Bhatia's book [2] and Horn and Johnson's book [7] are excellent sources for the background and preliminaries for the doubly stochastic matrices. Birkhoff's theorem provides an important representation of the set of doubly stochastic matrices which is related to many matrix inequalities. Denote the set of all N by N doubly stochastic matrices by $\mathscr{A}$ and the set of N by N permutation matrices by $\mathscr{P}$. Then we can state this result as

THEOREM 1.1. (Birkhoff)

$$\mathscr{A} = \operatorname{conv} \mathscr{P}.$$

REMARK 1.2. It is easy to verify that any $P \in \mathscr{P}$ is an extreme point of $\mathscr{A}$. Thus, $\mathscr{A} = \operatorname{conv} \mathscr{P}$ is equivalent to the original statement of the Birkhoff theorem: the set of extreme points of $\mathscr{A}$ is $\mathscr{P}$.

The purpose of this note is to give a short variational proof of this theorem. The existence of such a variational proof for the Birkhoff theorem is no surprising given connection of the doubly stochastic matrices with physics and economics. The proof here develops the method used in the proof of the Gordon alternatives by Borwein and Lewis in [3, Theorem 2.2.6]. It can also be used to provide a characterization of the level sets associated to the majorization. Variational methods have also been used in problems

Mathematics subject classification (2000): 15A42, 15A51, 58E30.

Key words and phrases: Variational method, doubly stochastic matrix, majorization, Birkhoff theorem.

The research was supported by NSF grant #0102496.

Paper MIA-7-34

related to doubly stochastic matrices by Borwein, Lewis and Nussbaum [4]. Traditional proofs of Birkhoff's theorem can be found, for example, in [2, 7]. A variational method usually refers to prove by arguing certain auxiliary function attains a minimum so that its derivative vanishes. After the discovery of general variational principles its mean has been extended to mathematical arguments using one of these variational principles (e.g. those in [6, 5]) and may involve nonsmooth functions. In particular, in this note we need the following approximate Fermat principle. This is a direct consequence of a finite dimensional version of the Borwein-Preiss smooth variational principle in [5]. We include a proof here for completeness.

THEOREM 1.3. (Approximate Fermat Principle)
Let V be a finite dimentional Banach space and let $f : V \to \mathbb{R}$ be a differentiable function. Suppose that f is bounded from below. Then, for any $\varepsilon > 0$, there exists $x \in V$ such that $\|f'(x)\| < \varepsilon$.

Proof. Choose $z \in V$ such that $f(z) < \inf_V f + \varepsilon/2$ and define $g(y) := f(y) + (\varepsilon/2)\|y - z\|^2$. Then g is continuous and coercive ($\lim_{\|y\|\to\infty} g(y) = +\infty$) so that it must attains its minimum, say, at $y = x$. It follows that

$$f(x) + (\varepsilon/2)\|x - z\|^2 \leqslant f(z) < \inf_V f + \varepsilon/2$$

and, therefore, $\|x - z\|^2 < 1$. On the other hand since g attains minimum at x, we have

$$f'(x) = -\frac{\varepsilon}{2}(\|\cdot\|^2)'(x - z).$$

Clearly, the norm of the right hand side is less than ε. □

2. Level sets related to majorization

We first give a variational proof of the characterization of the level sets corresponding to the *majorization*. This, in a simpler setting, illustrates the method we are going to use. For a vector $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$, we use $x^\downarrow$ to denote the vector derived from x by rearranging its components in a decreasing order. Recall that, for $x, y \in \mathbb{R}^N$, we say that x is *majorized* by y, denoted by $x \prec y$, provided that $\sum_{n=1}^N x_n = \sum_{n=1}^N y_n$ and $\sum_{n=1}^k x_n^\downarrow \leqslant \sum_{n=1}^k y_n^\downarrow$ for $k = 1, \ldots, N$. The level set for $y \in \mathbb{R}^N$ related to the majorization is defined by $l(y) := \{x \in \mathbb{R}^N : x \prec y\}$. We will show that $l(y) = \text{conv}\,\{Py : P \in \mathscr{P}\}$. To do so we first establish the following alternative characterization of the majorization.

LEMMA 2.1. *Let $x, y \in \mathbb{R}^N$. Then $x \prec y$ if and only if, for any $z \in \mathbb{R}^N$, $\langle z^\downarrow, x^\downarrow\rangle \leqslant \langle z^\downarrow, y^\downarrow\rangle$.*

Proof. Using Abel's formula we can write

$$\begin{aligned}\langle z^\downarrow, y^\downarrow\rangle - \langle z^\downarrow, x^\downarrow\rangle &= \langle z^\downarrow, y^\downarrow - x^\downarrow\rangle \\ &= \sum_{k=1}^{N-1}\left((z_k^\downarrow - z_{k+1}^\downarrow)\cdot\sum_{n=1}^{k}(y_n^\downarrow - x_n^\downarrow)\right) + z_N^\downarrow\sum_{n=1}^{N}(y_n^\downarrow - x_n^\downarrow).\end{aligned}$$

Now to see the necessity we observe that $x \prec y$ implies $\sum_{n=1}^{k}(y_n^{\downarrow} - x_n^{\downarrow}) \geqslant 0$ for $k = 1, \ldots, N-1$ and $\sum_{n=1}^{N}(y_n^{\downarrow} - x_n^{\downarrow}) = 0$. Thus, the last term in the right hand side of the previous equality is 0. Moreover, in the remaining sum each term is the product of two nonnegative factors and, therefore, is nonnegative. We now prove sufficiency. Suppose that, for any $z \in \mathbb{R}^N$,

$$0 \leqslant \langle z^{\downarrow}, y^{\downarrow}\rangle - \langle z^{\downarrow}, x^{\downarrow}\rangle = \sum_{k=1}^{N-1}\left((z_k^{\downarrow} - z_{k+1}^{\downarrow}) \cdot \sum_{n=1}^{k}(y_n^{\downarrow} - x_n^{\downarrow})\right) + z_N^{\downarrow}\sum_{n=1}^{N}(y_n^{\downarrow} - x_n^{\downarrow}).$$

Setting $z = \sum_{n=1}^{k} e_n$ for $k = 1, \ldots, N-1$ (where $\{e_n : n = 1, \ldots, N\}$ is the standard basis of $\mathbb{R}^N$) we have $\sum_{n=1}^{k} y_n^{\downarrow} \geqslant \sum_{n=1}^{k} x_n^{\downarrow}$, and setting $z = \pm\sum_{n=1}^{N} e_n$ we have $\sum_{n=1}^{N} y_n = \sum_{n=1}^{N} x_n$. □

THEOREM 2.2. (Representation of level sets related to the majorization) *Let $y \in \mathbb{R}^N$. Then*

$$l(y) = \operatorname{conv}\{Py : P \in \mathscr{P}\}.$$

Proof. It is clear that $l(y)$ is a convex set and $Py \prec y$ for any $P \in \mathscr{P}$. Thus, $\operatorname{conv}\{Py : P \in \mathscr{P}\} \subset l(y)$. To proof the reversed inequality, let $x \prec y$. For any $z \in \mathbb{R}^N$, choose $P \in \mathscr{P}$ such that

$$\langle z, Py\rangle = \langle z^{\downarrow}, y^{\downarrow}\rangle \geqslant \langle z^{\downarrow}, x^{\downarrow}\rangle \geqslant \langle z, x\rangle. \tag{1}$$

Observe that $\mathscr{P}$ is a finite set (with $N!$ elements to be precise). Thus, the function

$$g(z) := \ln\left(\sum_{P\in\mathscr{P}} \exp\langle z, Py - x\rangle\right)$$

is defined for all $z \in \mathbb{R}^N$, is differentiable and is bounded from below (by 0). By the approximate Fermat principle of Theorem 1.3 we can select a sequence $z_i \in \mathbb{R}^N$ such that

$$0 = \lim_{i\to\infty} f'(z_i) = \lim_{i\to\infty} \sum_{P\in\mathscr{P}} \lambda_P^i (Py - x), \tag{2}$$

where

$$\lambda_P^i = \frac{\exp\langle z_i, Py - x\rangle}{\sum_{P\in\mathscr{P}} \exp\langle z_i, Py - x\rangle}.$$

Clearly, $\lambda_P^i > 0$ and $\sum_{P\in\mathscr{P}} \lambda_P^i = 1$. Thus, taking a subsequence if necessary we may assume that, for each $P \in \mathscr{P}$, $\lim_{i\to\infty} \lambda_P^i = \lambda_P \geqslant 0$ and $\sum_{P\in\mathscr{P}} \lambda_P = 1$. Now taking limits as $i \to \infty$ in (2) we have

$$\sum_{P\in\mathscr{P}} \lambda_P (Py - x) = 0.$$

It follows that $x = \sum_{P\in\mathscr{P}} \lambda_P Py$, as was to be shown. □

Note that unlike in the Birkhoff theorem, for $P \in \mathscr{P}$, Py may not always be an extreme point of $l(y)$.

3. A variational proof of the Birkhoff Theorem

A similar argument can also be used to provide a variational proof of the Birkhoff theorem. We will need the following well known property of doubly stochastic matrices. Two different proofs of this fact are given in [1, 2].

LEMMA 3.1. *Let A be a doubly stochastic matrix. Then for some $P \in \mathscr{P}$, the entries in A corresponding to the $1's$ in P are all nonzero.*

We now turn to the proof of Birkhoff's theorem. It is an easy matter to verify that $\mathscr{A}$ is convex and $\mathscr{P} \subset \mathscr{A}$. Thus, conv $\mathscr{P} \subset \mathscr{A}$.

To prove the reversed inclusion, we view all $N \times N$ matrices as an (n^2 dimensional) Euclidean space $\mathscr{E}$ with inner product

$$\langle A, B\rangle = \operatorname{tr}(B^{\top}A) = \sum_{n,m=1}^{N} a_{nm}b_{nm}, \quad A, B \in \mathscr{E}.$$

We establish the following analogue of (1).

LEMMA 3.2. *Let $A \in \mathscr{A}$. Then for any $B \in \mathscr{E}$ there exists $P \in \mathscr{P}$ such that*

$$\langle B, A - P\rangle \geqslant 0.$$

Proof. We do an induction argument on the number of nonzero elements of A. Since every row and column of A sum to 1, A has at least N nonzero elements. If A has exactly N nonzero elements then they must all be one so that A itself is a permutation matrix and the lemma holds trivially. Suppose now that A has more than N nonzero elements. By Lemma 3.1 there exists $P \in \mathscr{P}$ such that the entries in A corresponding to the $1's$ in P are all nonzero. Let $t \in (0, 1)$ be the minimum of these N positive elements. Then we can verify that $A_1 = (A - tP)/(1 - t) \in \mathscr{A}$ and has at least one less nonzero elements than A. Thus, by the induction hypothesis there exists $Q \in \mathscr{P}$ such that

$$\langle B, A_1 - Q\rangle \geqslant 0.$$

It follows that $\langle B, A - tP - (1 - t)Q\rangle \geqslant 0$ and, therefore, at least one of $\langle B, A - P\rangle$ or $\langle B, A - Q\rangle$ is nonnegative. □

Now define a function f on $\mathscr{E}$ by

$$f(B) := \ln\left(\sum_{P\in\mathscr{P}} \exp\langle B, A - P\rangle\right).$$

Then f is defined for all $B \in \mathscr{E}$, is differentiable and is bounded from below by 0. By the approximate Fermat principle of Theorem 1.3 we can select a sequence $B_i \in \mathscr{E}$ such that

$$0 = \lim_{i\to\infty} f'(B_i) = \lim_{i\to\infty} \sum_{P\in\mathscr{P}} \lambda_P^i(A - P). \tag{3}$$

where

$$\lambda_P^i = \frac{\exp\langle B_i, A - P\rangle}{\sum_{P\in\mathscr{P}} \exp\langle B_i, A - P\rangle}.$$

Clearly, $\lambda_P^i > 0$ and $\sum_{P\in\mathscr{P}} \lambda_P^i = 1$. Thus, taking a subsequence if necessary we may assume that, for each $P \in \mathscr{P}$, $\lim_{i\to\infty} \lambda_P^i = \lambda_P \geqslant 0$ and $\sum_{P\in\mathscr{P}} \lambda_P = 1$. Now taking limits as $i \to \infty$ in (3) we have

$$\sum_{P\in\mathscr{P}} \lambda_P(A - P) = 0.$$

It follows that $A = \sum_{P\in\mathscr{P}} \lambda_P P$, as was to be shown. □

REFERENCES

[1] T. ANDO, *Majorization, doubly stochastic matrices and comparison of eigenvalues*, Linear Algebra and Its Applications, **118** (1989) 163–248.
[2] R. BHATIA, *Matrix Analysis*. Springer–Verlag, New York, 1997.
[3] J. M. BORWEIN AND A. S. LEWIS, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, Springer, New York, 2000.
[4] J. M. BORWEIN, A. S. LEWIS, AND R. NUSSBAUM, *Entropy minimization, DAD problems and doubly-stochastic kernels*, Journal of Functional Analysis, **123** (1994), 264–307.
[5] J. M. BORWEIN AND D. PREISS, *A smooth variational principle with applications to subdifferentiability and to differentiability of convex functions*, Trans. Amer. Math. Soc., **303** (1987), 517–527.
[6] I. EKELAND, *On the variational principle*, J. Math. Anal. Appl., **47** (1974), 324–353.
[7] R. A. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.

(Received October 10, 2003)

Qiji J. Zhu
Department of Mathematics
Western Michigan University
Kalamazoo, MI 49008
USA
e-mail: zhu@wmich.edu