JOURNAL OF APPROXIMATION THEORY **40**, 101–110 (1984)

# Orthogonal Projections Are Optimal Algorithms

CHARLES A. MICCHELLI

*IBM T. J. Watson Research Center,*
*Yorktown Heights, New York 10598 U.S.A.*

*Communicated by Oved Shisha*

Received December 25, 1982

Some results on worst case optimal algorithms and recent results of J. Traub, G. Wasilkowski, and H. Woźniakowski on average case optimal algorithms are unified. By the use of Housholder transformations it is shown that orthogonal projections onto the range of the adjoint of the information operator are, in a very general sense, optimal algorithms. This allows a unified presentation of average case optimal algorithms relative to Gaussian measures on infinite dimensional Hilbert spaces. The choice of optimal information is also discussed.

## BACKGROUND AND SETTING

As the title suggests, orthogonal projections are indeed optimal algorithms. For those who are familiar with the model of optimal estimation presented in [3], this is not surprising. The main contribution of this paper is to expand the generality in which this fact holds true.

Our motivation comes from the recent paper of Traub, Wasilkowski, and Woźniakowski [6]. They present some new optimality properties of orthogonal projections onto subspaces of *finite* dimensional Hilbert space $X$. (In their terminology, these projections are called *spline algorithms.*) In this paper, a substantial generalization of their result is given which is based on two observations. The first is the use of generalized Householder transformations and the second is the notion of unitary invariance. These ideas not only unify the theory of worst case and average case optimal algorithms but also apply to very general error criteria on spaces $X$ which need not be finite dimensional. They are applicable whenever the error criterion used satisfies properties related to the well-known class of unitary invariant norms [4]. Moreover, when a unitary invariant norm is used to measure the error, as in both worst case and average case models in Hilbert spaces, optimal information can also be obtained.

The specific questions we address below arise from a model of optimal estimation in normed linear spaces presented in [3]. To some degree, these

0021-9045/84 $3.00

ideas find their origin in what is sometimes called the *hypercircle inequality*. This is a method which gives sharp error bounds for estimating a linear functional of an unknown function when limited information is available about the function. The setting for the hypercircle inequality is a Hilbert space $X$ (not necessarily finite dimensional) with a bounded linear operator $I$ mapping $X$ onto $Y$. For simplicity, we assume that dim $Y$ is finite, but this is also not essential for the discussion in this section. Thus we may as well let $Y = R^n$ and for our convenience later we use the usual inner product $u \cdot v = \sum_{i=1}^{n} u_i v_i$ for vectors $u$, $v \in Y$. In the terminology of [3], $I$ is an *information operator* to be used in the following way: We wish to estimate $Lx$, where $x \in X$ and $L$ is a continuous linear functional on $X$ when the observation $Ix = y$ is made about $x$. Generally we require more information about $x$ to assess the error of any estimator. In our case, this takes the form, $x \in K$, where $K$ is the unit ball in $X$,

$$K = \{x : \|x\| \leqslant 1\}.$$

Thus the set of uncertainty in $x$ is the *hypercircle*

$$H = \{x : Ix = Ix_0, \|x\| \leqslant 1\}$$

determined by any $x_0 \in X$ such that $Ix_0 = y$. This hypercircle has a *Chebyshev center* given by $Qx_0 \in H$, where $Q$ is the orthogonal projection of $X$ onto $R(I^*)$. The hypercircle $H$ is taken by $L$ into an interval and the midpoint of this interval is $LQx_0$. Hence a best estimator for $Lx_0$ is $LQx_0$ and the error in estimation is given by

$$|Lx_0 - LQx_0| \leqslant (\|x_0\|^2 - \|Qx_0\|^2)^{1/2} \sup\{|Lw| : \|w\| \leqslant 1, Iw = 0\}$$

(see [2] for more details on these facts). The importance of this inequality rests on the fact that it holds universally for elements in the hypercircle $H$ and that $Qx_0$ *depends only on the data* $y$ (see below).

In [3], the following additional optimality property of $Q$ was observed. Let $U$ be any linear operator mapping $X$ into *any* normed linear space $Z$. (Generally, when there is more than one norm considered, even on the same space, we do not use any special notation to distinguish them.) We wish to estimate $Ux$, given $x \in K$ and the information $Ix$. We call *any* mapping $A : I(K) \to Z$ an *algorithm*. The algorithm $A$ yields an estimate $AIx$ for $Ux$. Since all we know is that $x \in K$ the error in estimating with this algorithm is

$$E(A) = \sup\{\|Ux - AIx\| : x \in K\}.$$

An optimal algorithm $A_0$ minimizes this error over *all* algorithms, that is,

$$E(A_0) = \min\{E(A) : A\}.$$

It was shown in [3] that $UQx = A_0 Ix$ is an optimal algorithm. Moreover, if

$$Ix = ((y_1, x), \dots, (y_n, x)), \qquad y_1, \dots, y_n \in X,$$

then

$$Q = \Sigma G_{ij}^{-1} y_i \otimes y_j,$$

where $G = (G_{ij})$ is the Gramian matrix $G_{ij} = (y_i, y_j)$ and $(y_i \otimes y_j)\, x = y_i(y_j, x)$. Thus the optimal algorithm is given explicitly by

$$A_0(v_1, \dots, v_n) = \Sigma v_i G_{ij}^{-1} U y_j. \tag{1}$$

It should be emphasized here that $Z$ does not have to be a Hilbert space.

In the following sections we provide improvements and refinements of not only these results but also of those contained in [5, 6].

## Average Case Optimality: The Finite Dimensional Case

We begin our discussion of average case optimality by requiring that both $X$ and $Z$ are *finite dimensional Hilbert spaces.* In this case, we measure the error by means of a pair $F = (f, d\mu)$, where $f$ is a strictly increasing convex and continuously differentiable function while $d\mu$ is Borel measure on $X$. For such a pair $F$, we define the $F$-error of $A$ to be

$$E_F(A) = \int_X f(\| Ux - AIx \|^2)\, d\mu(x).$$

We will always require that $d\mu$ is *unitarily invariant* relative to the norm on $X$. Thus

$$\int_X h(Rx)\, d\mu(x) = \int_X h(x)\, d\mu(x), \tag{2}$$

where $R$ is any isometry relative to the norm on $X$ and $h$ is any $\mu$-integrable function. An important example of such a measure is a Gaussian measure

$$d\mu(x) = \exp(-\|x\|^2)\, dx.$$

Also, in what follows we do not distinguish between algorithms $A_i$ such that $A_1 I = A_2 I$, a.e.

THEOREM 1. *Under the above hypotheses on the pair $F = (f, d\mu)$, $A_0$ is the unique F-optimal algorithm.*

*Proof.* Let $A$ be any algorithm, then

$$E_F(A) = \int_X f(\|(Ux - A_0 Ix) + (A_0 Ix - AIx)\|^2)\, d\mu(x).$$

Expanding the argument of $f$ and using its convexity we obtain

$$E_F(A) \geqslant \delta(\phi) + E_F(A^0) + \int_X f'(\|Ux - A_0 Ix\|^2)\|A_0 Ix - AIx\|^2\, d\mu(x),$$

where $A_0 Ix = Qx$, $Px = x - Qx$,

$$\delta(\phi) = 2 \int_X f'(\|UPx\|^2)(UPx, \phi(Ix))\, d\mu(x),$$

and

$$\phi(Ix) = A_0 Ix - AIx.$$

From this inequality the theorem will follow provided we can show $\delta(\phi) = 0$. To this end, we define $R = Q - P$. Note that $R$ is an isometry such that $IR = I$ and $PR = -P$. Hence using the unitary invariance of $d\mu(x)$ we have

$$\begin{aligned}\delta(\phi) &= 2 \int_X f'(\|UPRx\|^2)(UPRx, \phi(IRx))\, d\mu(x)\\ &= -\delta(\phi),\end{aligned}$$

which completes the proof of the theorem.

The case $f(t) = t$ of this theorem was proved by Traub, Wasilkowski and Woźniakowski in [6] by a more complicated argument.

It is worth observing that this result can be improved, provided we are willing to accept a data dependent hypothesis on $d\mu(x)$. Specifically, we have in mind measures of the form $g(x)\, d\mu(x)$, where $g(x)$ is invariant under the particular isometry used in the proof. For instance, if $g$ is the characteristic function of the preimage under $I$ of any subset of $Y$, the theorem remains valid for the measure $g d\mu$.

The generality of Theorem 1 also allows us to treat both *restricted* optimal algorithms and optimal algorithms for *inaccurate* information operators. Both of these possibilities were dealt with in [3] for the worst case model. In the first case, we suppose an algorithm is restricted to have its values in some fixed subspace $M$ of $Z$. Then the same proof shows that the unique restricted optimal algorithm is $P_M UQ$, where $P_M$ is the orthogonal projection of $Z$ onto $M$. In the second case, we suppose the exact information $Ix$ is not available, but rather an algorithm must use the inaccurate information $y = Ix - e$. The requirements of Theorem 1 leads us to suppose that the uncertainty in $x$ and the data error are measured together by a measure $d\mu$ which is unitarily invariant,

$$\iint f(\| Ux - A(Ix - e)\|^2)\, d\mu(x, e).$$

To apply Theorem 1 we should make the identification $U(x, y) = Ux$ and $I(x, y) = Ix - y$, for any $(x, y) \in XxY$. Then it is an easy matter to see that, when $XxY$ has its natural cross product norm, the $x$ component of the orthogonal projection $Q$ minimizes

$$\|x - u\|^2 + \| y - Iu\|^2$$

over all $u \in X$. From this observation it follows that the optimal algorithm is

$$A_0(v_1, \ldots, v_n) = \Sigma H_{ij}^{-1} v_j U y_i,$$

where $H_{ij} = G_{ij} + \delta_{ij}$ and $G_{ij} = (y_i, y_j)$, as before.

## Average Case Optimality: Infinite Dimensional Case

In the past section, we heavily relied upon the use of unitarily invariant measures. However, except for the existence of such a measure and its invariance under the family of isometries (2), the finite dimensionality of the underlying space was not used. We purposely presented our proofs in this manner to allow for their *immediate* extension to infinite dimensions. Of course, the existence of a unitarily invariant measure $\mu$ in this case becomes less apparent.

When $X$ is a separable Hilbert space and $S$ is a positive definite self-adjoint, trace class operator then there is a *Gaussian* measure $\mu$ whose covariance operator is $S$ (Prohorov; see [1, p. 29]). When $S$ is injective the range of $\sqrt{S}$ induces a Hilbert subspace of $X$, $X_0 = \sqrt{S(X)}$, with inner product

$$(\sqrt{S}x, \sqrt{S}y)_0 = (x, y).$$

Then $\mu$ is easily seen to be unitarily invariant relative to $X_0$ [1]. For us, this means that by restricting $I$ and $U$ to $X_0$ our previous results immediately extend to separable Hilbert spaces.

Since one of our goals is to show how worst case and average case analysis can be unified, we next present a worst case result.

## Worst Case Optimality

Let $f(s, t)$ be any real-valued nonnegative function defined for $s, t > 0$. We assume for every $t > 0$, $f(s, t)$ is nondecreasing and convex in $s$. We define the $f$-error of $A$ for estimating $U$ as

$$E_f(A) = \sup\{f(\|Ux - AIx\|, \|x\|) : x \in X\}. \tag{3}$$

Note that in (3) we take the supremum over *all* $x \in X$ rather than on $K$ as in the introduction. To specialize to that case we choose $f(s, t) = s/t$.

Theorem 2. *Suppose $f(s, t)$ is a real-valued nonnegative function defined for $s, t > 0$ such that $f(s, t)$ is convex and nondecreasing in $s$ for every $t$. Then $A_0$ defined by* (1) *is an $f$-optimal algorithm.*

*Proof.* The method we use to prove this result is different than that employed in Theorem 1. Here we follow the approach used in [3].

Let

$$e = \sup\{f(t\|Ux\|, t\|x\|) : x \in X, Ix = 0, t \geqslant 1\};$$

then we will show

$$e = E_f(A^0) = \min\{E_f(A) : A\}.$$

*Lower Bound*

Let $x$ be any element in $X$ with $Ix = 0$ and suppose $t \geqslant 1$. Given any algorithm $A$ we have

$$f(\|U(tx) - A(0)\|, \|tx\|) \leqslant E_f(A)$$

and

$$f(\|U(tx) + A(0)\|, \|tx\|) \leqslant E_f(A).$$

Since

$$t\|Ux\| \leqslant \tfrac{1}{2}\|U(tx) + A(0)\| + \tfrac{1}{2}\|U(tx) - A(0)\|,$$

it follows that

$$f(t\|Ux\|, t\|x\|) \leqslant E_f(A),$$

and so

$$e \leqslant \min\{E_f(A) : A\}.$$

*Upper Bound*

Suppose $A_0$ is defined by (1), then $UA_0Ix = UQx$ and

$$E_f(A_0) = \sup\{f(\|UPx\|, \|x\|) : x \in X\},$$

where $Px = x - Qx$. Since $IPx = 0$ and $\|x\| = t\|Px\|$, for some $t \geqslant 1$ it is clear that

$$E_f(A_0) \leqslant e,$$

and so the proof is complete.

## Main Point

We have seen above that both in the worst case and average case models of optimal estiation the orthogonal projection onto the range of the adjoint of the information operator leads to an optimal algorithm. The unifying feature of both of these results is embodied in the following observation.

Let $H$ be any nonnegative functional whose domain is all mapping from $X$ into $Z$. Suppose

(a) $H(-U) = H(U)$.

(b) $H(UR) = H(U)$ for all isometries $R$ relative to some inner product on $X$.

(c) $H(\frac{1}{2}(U + V)) \leqslant \max(H(U), H(V))$.

Then for any algorithm $A$, we use as before the isometry $R = Q - P$ and obtain

$$H(UP - AI) = H(UPR - AIR) = H(UP + AI).$$

Thus

$$H(UP) \leqslant H(UP - AI)$$

for all $A$. The particular cases considered before were

$$H(U) = \sup\{f(\|Ux\|, \|x\|) : x \in X\}, \qquad \text{worst case,}$$

$$= \int_X f(\|Ux\|^2)\, d\mu(x), \qquad \text{average case.}$$

## Optimal Information

We now turn our attention to an *optimal* choice for the information operator $I$. This has been done for the worst case model in [3]. Let us now consider the problem for the average case model.

## AVERAGE CASE OPTIMAL INFORMATION: THE FINITE DIMENSIONAL CASE

In this section, we restrict ourselves to $f(t) = t^p$, $p > 1$, and as before $d\mu$ is a unitarily invariant measure. For convenience we refer to the corresponding $F$-error of an algorithm as the $p$-error.

Suppose now that $Z$ is *also* a Hilbert space. For any bounded linear operator $T: X \to Z$ we define

$$|T| = \left(\int_X |(Tx, Tx)|^p \, d\mu(x)\right)^{1/2p}.$$

It is easily seen that $|\cdot|$ is a unitarily invariant norm in the sense of [4], i.e., $|ST| = |TR| = |T|$ for any isometries $S, R$. Moreover, it is clear that

$$\begin{aligned} E_F(A_0) &= |UP|^{2p} \\ &= |U - UQ|^{2p}, \end{aligned}$$

and

$$\min\{E_F(A_0; I): I\} \geqslant \min\{|U - T|^{2p} : \dim \text{range } T \leqslant n\}.$$

According to [4], the lower bound can be evaluated from the singular value decomposition of $U$. Thus expressing $U^*U$ as

$$U^*U = \sum_{i=1}^{m} \sigma_i y_i^0 \otimes y_i^0, \tag{4}$$

where $y_1^0, ..., y_m^0$ are the orthonormal column eigenvectors of $U^*U$ with corresponding eigenvalues $\sigma_1, ..., \sigma_m$ ordered so that $\sigma_1 \geqslant \cdots \geqslant \sigma_m$. (These are called the *singular values* of $U$.) Then the lower bound above is achieved by the operator

$$T_{opt} = \sum_{i=1}^{n} Uy_i^0 \otimes y_i^0,$$

which gives us the inequality

$$\min\{E_F(A_0, I): I\} \geqslant |U - T_{opt}|^{2p}.$$

Since $T_{\text{opt}} = UQ_{\text{opt}}$, where

$$Q_{\text{opt}} = \sum_{i=1}^{n} y_i^0 \otimes y_i^0,$$

we finally obtain $\min_I\{E_F(A_0; I)\} = E_F(A_{\text{opt}}, I_{\text{opt}})$

for

$$A_{\mathrm{opt}}(v_1,\ldots,v_n)=\sum_{j-1}^{n} v_j U y_j^0 \tag{5}$$

and

$$I_{\mathrm{opt}}(x)=((y_1^0,x),\ldots,(y_n^0,x)).$$

We state these observations below as

THEOREM 3. *Suppose* $U^*Uy_i^0=\sigma_i y_i^0$, $\sigma_1\geqslant\cdots\geqslant\sigma_m\geqslant 0$ *and* $(y_i,y_j)=\delta_{ij}$. *Then an optimal linear information operator for estimating Ux relative to the p-error*

$$E_p(A)=\int_X \|Ux-A(Ix)\|^{2p}\,d\mu(x)$$

*is given by* (6) *and the corresponding unique optimal algorithm by* (5).

The optimality of $I_{\mathrm{opt}}$ was known in the worst case model when $f(t,s)=t$. This is a standard consequence of the theory of $n$-widths in Hilbert spaces and is explained in [3]. In the context of average case optimality the above remarks suggest a notion of *average n-width.* Suppose $X$ is a normed linear space, and $d\mu(x)$ a Borel measure defined on $X$. We define the $p$th average $n$-width of the set $UK$ relative to $d\mu$ as

$$d_n^a(UK)=\inf\left\{\left(\int_X \operatorname{dist}(Ux,X_n)^p\,d\mu(x)\right)^{1/p} : X_n\subseteq X, \dim X_n=n\right\},$$

where

$$\operatorname{dist}(x,X_n)=\inf\{\|x-y\| : y\in X_n\}.$$

When $p\geqslant 2$, $X$ is a Hilbert space and $d\mu$ a unitarily invariant measure, our previous remarks can be used to easily identify $d_n^a$ because

$$\operatorname{dist}(Ux,X_n)=\|U*x-U*Qx\|,$$

where $Q$ is the orthogonal projection of $X$ onto $X_n$. It would be interesting to determine $d_n^a(UK)$ in other cases.

## REFERENCES

1. HUI-HSUING KUO, *in* "Gaussian Measures in Banach Spaces," Lecture Notes in Mathematics, No. 463, Springer-Verlag, Berlin, 1975.

2. A. A. Melkman and C. A. Micchelli, Optimal estimation of linear operators in Hilbert spaces from inaccurate date, *SIAM J. Numer. Anal.* **16** (1979), 87–105.
3. C. A. Micchelli and T. J. Rivlin, A survey of optimal recovery, *in* "Optimal Estimation in Approximation Theory" (C. A. Micchelli and T. J. Rivlin, Eds.), Plenum, New York, 1976.
4. L. Mirsky, Symmetric gauge functions and unitarily invariant norms, *Quart. J. Math. Oxford Ser.* (2) **11** (1960), 50–59.
5. J. Traub and H. Woźniakowski, "A General Theory of Optimal Algorithms," Academic Press, New York, 1980.
6. J. Traub, G. Wasilkowski, and H. Woźniakowski, "Average Case Optimality for Linear Problems," Department of Computer Science Report, Columbia University, 1981.