Title

A linear surrogate for optimising functions of an orthogonal matrix with applications in wave function theory

Journal

Molecular Physics, 121(9-10)

ISSN

0026-8976

Authors

Wang, Zhenling

Head-Gordon, Martin

Publication Date

2023-05-19

DOI

10.1080/00268976.2022.2118185

A Linear Surrogate for Optimizing Functions of an Orthogonal Matrix with Applications in Wave Function Theory

Zhenling Wang and Martin Head-Gordon

Department of Chemistry, University of California, and,

Chemical Sciences Division, Lawrence Berkeley National Laboratory,

Berkeley CA 94720, USA

Abstract

The technique of surrogate optimization is to use a simpler function to approximate a complex function that is time-consuming to evaluate. We show that the maximum of a special type of surrogate function $f(U) = \mathrm{Tr}(AU), U \in O(n)$ is at $A^T(AA^T)^{1/2}$, and that there is one and only one local maximum both in $SO(n)$ and $O(n) - SO(n)$. This function $f(U)$ has been found to be useful in various aspects of electronic structure theory, including proving the Carlson-Keller theorem, and localizing orbitals. As one other example, we apply it here to optimize the ground state of molecules using the Generalized Valence Bond wavefunction.

Keywords

surrogate optimization, orthogonal matrix, orbital localization, Generalized Valence Bond theory

1. Introduction

Optimization problems are ubiquitous in quantum chemistry, and sometimes the objective function is hard to evaluate. For example, in geometry optimization, we are finding local minima (or saddle points) on molecular potential energy surfaces, and in self-consistent field (SCF) orbital optimization, we are trying to reach the minimum on a Grassman manifold. Both problems are non-trivial due to the computational cost of evaluating the objective function and its gradients at a large number of trial points. Therefore, we can achieve a considerable speedup if we construct an approximate function that is fast to evaluate – this is the idea of surrogate optimization. Recently, this idea has attracted some interest in chemistry, with research emphasis particularly on using machine learning to do geometry optimization[1–4]. There are also applications in experimental design[5], predicting molecular properties[6] and calculating the electronic structure[7,8].

It is uncommon to see surrogate optimization discussed in electronic structure theory. There is one common example: the Roothaan step in SCF optimization[9] updates the density matrix by finding the global minimum (or a saddle point in exotic cases[10–12]) of a surrogate function, $E = \mathrm{Tr}(PF)$, where F is the current Fock matrix (taken as constant) in an orthogonal basis. This surrogate function is directly optimized by diagonalizing F, yielding the density matrix as $P = \sum_{j=1}^{O} C_j C_j^T$ where C_j are the orthonormal eigenvectors corresponding to the O (number of electrons) lowest eigenvalues: $FC_j = C_j \varepsilon_j$. Apart from this classic example, and the two machine-learning-relevant papers mentioned above, there is only one highly developed idea – the second-order convergence method[13]. For instance, in multi-configuration SCF (MCSCF)[13–16] and generalized valence bond (GVB) theory[17,18], we want to minimize the system energy E, which is a function of an orthogonal matrix $E(U)$. U can be written as the matrix exponential of an antisymmetric matrix $U = \exp(\Delta)$, and second-order convergence suggests expand E either to the second order of Δ and minimize the surrogate analytically or to the second order of $T = U - I$ and minimize it iteratively. In this paper, however, we will find the exact extremum point of a U-linearized surrogate function $\mathrm{Tr}(AU)$, without the need to further approximate U by a polynomial or to iteratively solve a set of non-linear equations.

Based on the following lemma, we can find the exact extremum of $f(U) = \mathrm{Tr}(AU)$.

Lemma 1 Given $A \in \mathbb{R}^{n\times n}$ an invertible matrix and $f: SO(n) \to \mathbb{R}$ defined by

$$f(U) = \mathrm{Tr}(AU),$$

then f has one and only one local maximum. Following standard notation[19], $SO(n)$ refers to the special orthogonal group of rotation matrices with determinant equal to 1, in contrast to the more general orthogonal group, $O(n)$, which also includes e.g. reflection matrices. Subotnik and co-workers attempted to prove it[20], and claimed that the maximum point of f is

$$U_1 = A^T(AA^T)^{-1/2}.$$

However, this is incorrect, consider

$$A = \begin{pmatrix} 1 & \\ & -1 \end{pmatrix},$$

The resulting U_1 will be

$$U_1 = \begin{pmatrix} 1 & \\ & -1 \end{pmatrix} \notin SO(n),$$

which is clearly contradictory. In this paper, we will give an accurate statement and complete proof of this lemma in Section 2. Then, several applications are considered in Section 3, including orbital orthogonalization, where another proof of the Carlson-Keller theorem[21] is given, orbital localization, where we ensure that the hidden conjecture made by Subotnik et al. is actually reasonable, and finally orbital optimization in GVB theory. This lemma, as pointed out by an anonymous reviewer, is equivalent to the "orthogonal Procrustes problem"[22] used in many other fields, including psychometrics[23], crystallography[24,25], computer science[26], bioinformatics[27], and is a special case of the Wahba's problem[28]. But as we have mentioned before, this type of surrogate approach is rare in electronic structure theory. We hope this paper might attract some interest in proposing and analyzing new surrogate optimization models in this field.

2. Lemma and its Proof

2.1. Global Maximum

Subotnik et al.'s maximum point is actually correct in $O(n)$. That the global maximum is indeed U_1 is not difficult to prove.

Lemma 2 Given $A \in \mathbb{R}^{n\times n}$ an invertible matrix and $f: O(n) \to \mathbb{R}$ defined by

$$f(U) = \mathrm{Tr}(AU),$$

then the unique global maximum of f is at U_1.

Proof

We perform singular value decomposition (SVD) on A, to obtain

$$A = F\Sigma G^T,$$

where $F, G \in O(n)$ and Σ is diagonal with all diagonal values (singular values of A) positive. Now,

$$\mathrm{Tr}(AU) = \mathrm{Tr}(F\Sigma G^T U) = \mathrm{Tr}[\Sigma(G^T UF)],$$

where $G^T UF \in O(n)$. As $G^T UF$ is orthogonal, its diagonal elements must not have absolute values larger than 1, so

$$\mathrm{Tr}[\Sigma(G^T UF)] \leq \mathrm{Tr}(\Sigma),$$

and the equality is achieved only when

$$G^T UF = I_n \Leftrightarrow U = GF^T;$$

here I_n is the identity matrix. Actually, as $AA^T = F\Sigma^2 F^T$, we know that

$$(AA^T)^{-1/2} = F\Sigma^{-1}F^T,$$

so

$$U_1 = A^T(AA^T)^{-1/2} = GF^T.$$

□

For the case that f is restricted on $SO(n)$, the location of the maximum depends on whether $\det A$ is positive or negative.

Lemma 2' Given $A \in \mathbb{R}^{n\times n}$ is an invertible matrix with SVD $A = F\Sigma G^T$ (Σ is sorted to have larger values on the upper left part of the matrix) and $f: SO(n) \to \mathbb{R}$ defined by

$$f(U) = \mathrm{Tr}(AU),$$

then the unique global maximum of f is

(a) at $GF^T = U_1$ if $\det A > 0$;

(b) at $GLF^T = U_{-1}$ if $\det A < 0$ and the smallest singular value of A is non-degenerate; here $L = \mathrm{diag}(1, \cdots, 1, -1)$.

Proof

(a) follows immediately from *Lemma 2* as $G^T UF \in SO(n)$, we only need to prove (b).

We still write

$$\mathrm{Tr}(AU) = \mathrm{Tr}[\Sigma(G^T UF)] = \mathrm{Tr}(\Sigma P),$$

and this time $P \in O(n) - SO(n)$. We denote the i^{th} diagonal element of matrix Σ as σ_{ii}, and clearly

$$\begin{aligned}\mathrm{Tr}(\Sigma P) &= \sum_i \sigma_{ii} P_{ii} = \sum_i (\sigma_{ii} - \sigma_{nn}) P_{ii} + \sigma_{nn}\,\mathrm{Tr}(P) \\ &\leq \sum_i (\sigma_{ii} - \sigma_{nn}) \cdot 1 + \sigma_{nn} \cdot (n-2) = \sum_i \sigma_{ii} - 2\sigma_{nn} \\ &= \mathrm{Tr}(\Sigma) - 2\sigma_{nn},\end{aligned}$$

where $\mathrm{Tr}(P)$, being the sum of all eigenvalues of P, cannot exceed $n-2$ as $\det P = -1$. The equality is reached, when $\sigma_{nn} \neq \sigma_{n-1,n-1}$, only in the case that

$$P_{11} = \cdots P_{n-1,n-1} = 1,$$

and

$$\mathrm{Tr}(P) = n - 2,$$

meaning

$$P = L = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & -1 \end{pmatrix}.$$

Also, if $\sigma_{nn} = \sigma_{n-1,n-1}$, the maximum point is not unique, as

$$P' = \mathrm{diag}(1, \cdots, 1, -1, 1)$$

satisfies $\mathrm{Tr}(\Sigma P') = \mathrm{Tr}(\Sigma) - 2\sigma_{nn}$ as well. Thus, if conditions in (b) are satisfied, the maximum point of f is

$$G^T UF = L \Leftrightarrow U = GLF^T = U_{-1}.$$

□

2.2. Local Maximum when $\det A > 0$

To find any **local** maximum, we first consider the case that $\det A > 0$. In this case, because $G^T SO(n) F = SO(n)$, the problem is equivalent to finding the local maximum of

$$f(U) = \mathrm{Tr}(\Sigma U),$$

where Σ is a positive sorted diagonal matrix, i.e.,

$$\Sigma = \begin{pmatrix} \sigma_1 I_{k_1} & & & \\ & \sigma_2 I_{k_2} & & \\ & & \ddots & \\ & & & \sigma_m I_{k_m} \end{pmatrix}, \tag{1}$$

where $k_1, k_2, \cdots, k_m$ are the multiplicity of singular values $\sigma_1 > \sigma_2 > \cdots > \sigma_m > 0$, and I_k is the identity matrix of dimension k.

Any local maximum must satisfy two conditions: first it must be a stationary point, and second its function value must not be smaller than that of a point in its neighborhood. The condition for a stationary point is presented in the *Lemma 3* below.

Lemma 3 Given a positive sorted diagonal matrix Σ (the definition of such matrices is given above), the stationary points of function $f: SO(n) \to \mathbb{R}$ defined by $f(U) = \mathrm{Tr}(\Sigma U)$ are

$$U = \begin{pmatrix} U_{k_1} & & & \\ & U_{k_2} & & \\ & & \ddots & \\ & & & U_{k_m} \end{pmatrix}, \tag{2}$$

where $U_{k_1}, U_{k_2}, \cdots, U_{k_m}$ are symmetric orthogonal matrices of dimensions $k_1, k_2, \cdots, k_m$.

Proof

At a certain stationary point candidate U_0, we consider a shifted function $f_0: SO(n) \to \mathbb{R}$ and

$$f_0(U) = \mathrm{Tr}(\Sigma U_0 U),$$

and we write $U = \exp(\Delta)$ where Δ is antisymmetric. Differentiating with respect to elements of Δ at the point $\Delta = 0$ gives

$$\left.\frac{\partial f_0}{\partial \Delta_{pq}}\right|_{\Delta=0} = \sum_{i,j} (\Sigma U_0)_{ij} \left.\frac{\partial U_{ji}}{\partial \Delta_{pq}}\right|_{\Delta=0}$$

$$= \sum_{i,j} (\Sigma U_0)_{ij} \left(\delta_{jp}\delta_{iq} - \delta_{jq}\delta_{ip} \right)$$
$$= (\Sigma U_0)_{qp} - (\Sigma U_0)_{pq}.$$

Therefore, stationary point U_0 must have $\Sigma U_0 = (\Sigma U_0)^T$. We claim that every U satisfying $\Sigma U = (\Sigma U)^T$ has the form in Equation (2), and if that is verified, *Lemma 3* is proved.

Let

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{nn} \end{pmatrix},$$

and $\Sigma = \text{diag}(\sigma_{11}, \cdots, \sigma_{nn})$ where $\sigma_{11} \geq \sigma_{22} \geq \cdots \geq \sigma_{nn} > 0$. Thus,

$$\Sigma U = \begin{pmatrix} \sigma_{11}u_{11} & \sigma_{11}u_{12} & \cdots & \sigma_{11}u_{1n} \\ \sigma_{22}u_{21} & \sigma_{22}u_{22} & \cdots & \sigma_{22}u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{nn}u_{n1} & \sigma_{nn}u_{n2} & \cdots & \sigma_{nn}u_{nn} \end{pmatrix},$$

and as ΣU is symmetric, the squared norm of its first row must be equal to that of its first column, i.e.,

$$\sigma_{11}^2(u_{11}^2 + u_{12}^2 + \cdots + u_{1n}^2) = \sigma_{11}^2 u_{11}^2 + \sigma_{22}^2 u_{21}^2 + \cdots + \sigma_{nn}^2 u_{n1}^2.$$

However, U is orthogonal, so $u_{11}^2 + u_{12}^2 + \cdots + u_{1n}^2 = u_{11}^2 + u_{21}^2 + \cdots + u_{n1}^2 = 1$, which means

$$\sigma_{11}^2 = \sigma_{11}^2 u_{11}^2 + \sigma_{22}^2 u_{21}^2 + \cdots + \sigma_{nn}^2 u_{n1}^2 \leq \sigma_{11}^2(u_{11}^2 + u_{21}^2 + \cdots + u_{n1}^2) = \sigma_{11}^2,$$

so, for $i = 2, \cdots, n$, either $\sigma_{ii}^2 = \sigma_{11}^2$ or $u_{i1} = u_{1i} = 0$. Considering that Σ has the form in Equation (1), we know that U must now be of the form

$$U = \begin{pmatrix} U_{k_1} & & & \\ & * & * & * \\ & * & \ddots & * \\ & * & * & * \end{pmatrix},$$

where U_{k_1} is a $k_1 \times k_1$ matrix and $*$ stands for a part that is not yet determined. Because the first k_1 rows of U are orthonormal, U_{k_1} is orthogonal. Also, as ΣU is symmetric, $\sigma_1 U_{k_1}$ is symmetric, and thus U_{k_1} is symmetric.

Similarly, we can apply the same argument to the $(k_1 + 1)^{\text{th}}$ row and $(k_1 + 1)^{\text{th}}$ column, getting similar results; finally, we will arrive at

$$U = \begin{pmatrix} U_{k_1} & & & \\ & U_{k_2} & & \\ & & \ddots & \\ & & & U_{k_m} \end{pmatrix},$$

where $U_{k_1}, U_{k_2}, \cdots, U_{k_m}$ are symmetric orthogonal matrices but their determinants are not necessarily positive.

□

We know that the function value of a local maximum cannot be smaller than that of a point in its neighborhood, and we will test this requirement at all stationary points; the result is *Lemma 4*.

Lemma 4 Given the matrix Σ and function f as in *Lemma 3*, at all stationary points, only $U = I_n$ can be a local maximum, and as $U = I_n$ is the global maximum point (*Lemma 2'*), it is indeed a local maximum.

Proof

Firstly, U can be diagonalized:

$$U = W\Lambda W^T,$$

where W is a block diagonal orthogonal matrix $W = \text{diag}(W_{k_1}, \cdots, W_{k_m})$ and Λ is a diagonal matrix with elements 1 or -1; we write Λ in a block diagonal form $\Lambda = \text{diag}(\Lambda_{k_1}, \cdots, \Lambda_{k_m})$. Notice that $W_{k_1}, \cdots, W_{k_m}$ are all orthogonal matrices, and $U_{k_i} = W_{k_i}\Lambda_{k_i}W_{k_i}^T$ for all i.

We will prove this *lemma* in two steps: considering U being a stationary point and a local maximum of f, we will first show that each block of U cannot have two -1 eigenvalues, and then we will show that all -1 eigenvalues are in the same block of U. With these two claims, U can have only one -1 eigenvalue at most; considering that $\det U > 0$, all eigenvalues must be 1, and thus $U = WW^T = I_n$.

To see that each block of U cannot have two negative eigenvalues, without losing generality we can assume that the first two eigenvalues of the first block U_{k_1} are -1, i.e.,

$$\Lambda_{k_1} = \text{diag}(-1, -1, \pm 1, \cdots, \pm 1),$$

where ± 1 stands for elements that can be either 1 or -1. We can build, for $\theta \in [0, 2\pi)$,

$$\Lambda'_{k_1} = \text{diag}\left(\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \pm 1, \cdots, \pm 1\right),$$

$\Lambda' = \text{diag}(\Lambda'_{k_1}, \Lambda_{k_2}, \cdots, \Lambda_{k_m})$, and $U' = W\Lambda' W^T$. It is clear that for any neighborhood of U in $SO(n)$, there exists a θ that is close but not equal to π satisfying that U' is in that neighborhood. But we see that

$$\begin{aligned} \text{Tr}(\Sigma U') - \text{Tr}(\Sigma U) &= \text{Tr}(\Sigma W(\Lambda' - \Lambda)W^T) \\ &= \sigma_1 \text{Tr}\left(W_{k_1}(\Lambda'_{k_1} - \Lambda_{k_1})W_{k_1}^T\right) \\ &= \sigma_1 \text{Tr}\left((\Lambda'_{k_1} - \Lambda_{k_1})W_{k_1}W_{k_1}^T\right) \\ &= \sigma_1 \text{Tr}(\Lambda'_{k_1} - \Lambda_{k_1}) = 2\sigma_1(\cos\theta + 1) > 0, \end{aligned}$$

so U cannot be a local maximum.

In the case that two different blocks have -1 eigenvalues, we can assume that the first eigenvalues of the first two blocks U_{k_1}, U_{k_2} are -1. Λ' can be built similarly as before, for a $\theta \in [0, 2\pi)$ that is close but not equal to π,

$$\Lambda' = \text{diag}\left(\left(\begin{pmatrix} \cos\theta & & \sin\theta & \\ & \pm 1 & & \\ & & \ddots & \\ -\sin\theta & & \cos\theta & \\ & & & \pm 1 \\ & & & & \ddots \end{pmatrix}_{(k_1+k_2)\times(k_1+k_2)}, \Lambda_{k_3}, \cdots, \Lambda_{k_m}\right),$$

and it can be ensured that such θ exists for any neighborhood of U such that matrix $U' = W\Lambda' W^T$ is in that neighborhood. We can calculate that

$$\begin{aligned} &\text{Tr}(\Sigma U') - \text{Tr}(\Sigma U) \\ &= \text{Tr}(\Sigma W(\Lambda' - \Lambda)W^T) \\ &= \text{Tr}\left(\begin{pmatrix} \sigma_1 I_{k_1} & \\ & \sigma_2 I_{k_2} \end{pmatrix}\begin{pmatrix} W_{k_1} & \\ & W_{k_2} \end{pmatrix}\begin{pmatrix} \cos\theta + 1 & & \sin\theta & \\ & 0 & & \\ & & \ddots & \\ -\sin\theta & & \cos\theta + 1 & \\ & & & 0 \\ & & & & \ddots \end{pmatrix}\begin{pmatrix} W_{k_1}^T & \\ & W_{k_2}^T \end{pmatrix}\right) \end{aligned}$$

$$= \left(\sigma_1 \sum_{i=1}^{k_1} w_{k_1,1i}^2 + \sigma_2 \sum_{i=1}^{k_2} w_{k_2,1i}^2 \right) (\cos\theta + 1) = (\sigma_1 + \sigma_2)(\cos\theta + 1) > 0,$$

where $w_{k_1,1i}$ is an element in the first row of W_{k_1} and $w_{k_2,1i}$ is defined similarly; the last equality follows from the fact that W_{k_1}, W_{k_2} are orthogonal. The fact that $\mathrm{Tr}(\Sigma U') > \mathrm{Tr}(\Sigma U)$ means U cannot be a local maximum, so all -1 eigenvalues must be in the same block, and the proof is thus complete.

□

With *Lemma 3* and *Lemma 4*, we know that *Lemma 1* is true when $\det A > 0$, and the only local maximum is $U_1 = A^T(AA^T)^{-1/2}$.

2.3. Local Maximum when $\det A < 0$

When $\det A < 0$, $G^T SO(n) F \neq SO(n)$, but

$$\mathrm{Tr}(AU) = \mathrm{Tr}(F\Sigma G^T U) = \mathrm{Tr}(\Sigma L L G^T U F),$$

where the matrix $L = \mathrm{diag}(1, \cdots, 1, -1)$, and now $LG^T SO(n) F = SO(n)$, so by defining $\Sigma' = \Sigma L$, it is sufficient to prove that function $f(U) = \mathrm{Tr}(\Sigma' U)$ has only one local maximum in $SO(n)$; here

$$\Sigma' = \begin{pmatrix} \sigma_1 I_{k_1} & & & \\ & \ddots & & \\ & & \sigma_{m-1} I_{k_{m-1}} & \\ & & & -\sigma_m \end{pmatrix},$$

and we need σ_m to be non-degenerate to ensure that the global maximum is unique.

The proof of *Lemma 3* does not require Σ to be positive, it only requires elements with larger **absolute** values to be in the upper left part of the matrix, which is also the case of Σ' – nothing needs to be done there. The first part of *Lemma 4*, showing that two -1 eigenvalues cannot be in the same block, requires the relevant σ to be positive; but the block in Σ' that has negative σ is only one-dimensional, meaning that finding two -1 eigenvalues is naturally impossible. The second part of *Lemma 4* is showing all -1 eigenvalues are in the same block. The proof only requires the sum of σ in the relevant

blocks to be positive, but $\sigma_1 > \sigma_2 > \cdots > \sigma_m > 0$, so that requirement is satisfied. Therefore, *Lemma 3* and *Lemma 4* are applicable to $f(U) = \mathrm{Tr}(\Sigma' U)$, giving the only local maximum at $U = I_n$, which means that *Lemma 1* is also true in this case and that the local maximum is

$$LG^T UF = I_n \Leftrightarrow U = GLF^T = U_{-1}.$$

Finally, the complete form of *Lemma 1* is summarized and relabeled below,

Lemma 5 Given $A \in \mathbb{R}^{n\times n}$ an invertible matrix with its SVD $A = F\Sigma G^T$ (Σ is sorted to have larger values on the upper left part of the matrix) and function $f: SO(n) \to \mathbb{R}$ defined by

$$f(U) = \mathrm{Tr}(AU),$$

then f has one and only one local maximum. The maximum point is

(a) at $GF^T = A^T(AA^T)^{-1/2} = U_1$ if $\det A > 0$;

(b) at $GLF^T = U_{-1}$ if $\det A < 0$ and the smallest singular value of A is non-degenerate, where $L = \mathrm{diag}(1, \cdots, 1, -1)$.

3. Applications

3.1. Löwdin Symmetric Orthogonalization

Löwdin symmetric orthogonalization[29] is a method to orthogonalize orbitals. For a set of n linearly independent orbitals $\{\phi_i\}_{i=1}^n$ with their coefficient matrix C_0 and overlap matrix S_0, symmetric orthogonalization

$$C_1 = C_0 S_0^{-1/2}$$

orthogonalizes the set. The coefficient matrix C of any orthogonalized set $\{\phi_i'\}_{i=1}^n$ can be conveniently written as

$$C = C_0 S_0^{-1/2} U,$$

where U is an arbitrary matrix in $O(n)$. The Carlson-Keller theorem[21,30–32] claims that the symmetric orthogonalized orbital set resembles the original set most, i.e., $U = I$ minimizes the function

$$f(U) = \sum_{i=1}^{n} \langle \phi_i - \phi_i' | \phi_i - \phi_i' \rangle = 2n - 2\sum_{i=1}^{n} \langle \phi_i | \phi_i' \rangle = 2n - 2\mathrm{Tr}\left(S_0^{1/2} U\right),$$

and we know directly from *Lemma 5* that the global minimum point is indeed $U = I$, and that $U = I$ is also the only local minimum in $SO(n)$.

This result can be directly generalized to weighted orthogonalization[33]. In that case, the function to be minimized is

$$g(U) = \sum_{i=1}^{n} w_i \langle \phi_i - \phi_i' | \phi_i - \phi_i' \rangle = 2\sum_{i=1}^{n} w_i - 2\sum_{i=1}^{n} w_i \langle \phi_i | \phi_i' \rangle = 2\sum_{i=1}^{n} w_i - 2\mathrm{Tr}\left(W S_0^{1/2} U\right),$$

here the positive weights are $\{w_i\}_{i=1}^{n}$ and the matrix W is a diagonal matrix

$$W = \mathrm{diag}(w_1, w_2, \cdots, w_n).$$

From *Lemma 5*, we know that the global minimum point of g is

$$U = S_0^{1/2} W (W S_0 W)^{-1/2},$$

so the matrix of the weighted orthogonalized orbital set is

$$C = C_0 W (W S_0 W)^{-1/2},$$

which is the same as the classic result, and we also know that U is the only local minimum in $SO(n)$. The fact that the local minimum is unique justifies iterative algorithms designed for these processes[34].

3.2. Surrogate Optimization in Orbital Localization

Orbital localization techniques are crucial to understanding chemical bonding, and also underpin most of the linear scaling algorithms in quantum chemistry. Three important examples are Boys[35,36], Pipek-Mezey[37], and Edmiston-Ruedenberg[38] localization, which maximize the following functionals of $U \in O(n)$ given a set of orthonormalized orbitals $\{\phi_i\}_{i=1}^{n}$.

$$\text{Boys: } f_B(U) = \sum_{i=1}^{n} |\langle \phi_i | \boldsymbol{r} | \phi_i \rangle|^2,$$

$$\text{Pipek-Mezey: } f_{PM}(U) = \sum_{i=1}^{n} \sum_{A=1}^{N_A} |\langle \phi_i | P_A | \phi_i \rangle|^2,$$

$$\text{Edmiston-Ruedenberg: } f_{ER}(U) = \sum_{i=1}^{n} \langle \phi_i \phi_i | \frac{1}{r} | \phi_i \phi_i \rangle,$$

where in f_{PM}, N_A is the number of atoms and P_A is the projection operator onto the space spanned by the atomic orbitals of atom A. These functionals are not linear with respect to U, so *Lemma 5* cannot be utilized directly. However, we can easily write surrogate functions and hope that a series of surrogate optimizations will converge to the global maximum. In other words, we write surrogate functions

$$\text{Boys: } f_B^{'}(U) = \sum_{i=1}^{n} \langle \phi_i | \boldsymbol{r} | \phi_{i,0} \rangle \cdot \langle \phi_{i,0} | \boldsymbol{r} | \phi_{i,0} \rangle,$$

$$\text{Pipek-Mezey: } f_{PM}^{'}(U) = \sum_{i=1}^{n} \sum_{A=1}^{N_A} \langle \phi_i | P_A | \phi_{i,0} \rangle \langle \phi_{i,0} | P_A | \phi_{i,0} \rangle,$$

$$\text{Edmiston-Ruedenberg: } f_{ER}^{'}(U) = \sum_{i=1}^{n} \langle \phi_i \phi_{i,0} | \frac{1}{r} | \phi_{i,0} \phi_{i,0} \rangle,$$

where $\{\phi_{i,0}\}_{i=1}^{n}$ is a fixed set of orbitals. All three surrogate functionals can be recast into the form of $\mathrm{Tr}(AU)$; for example, $f_B^{'}(U)$ can be written as

$$f_B^{'}(U) = \sum_{i=1}^{n} \langle \phi_i | \boldsymbol{r} | \phi_{i,0} \rangle \cdot \langle \phi_{i,0} | \boldsymbol{r} | \phi_{i,0} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} U_{ji} \langle \phi_{j,0} | \boldsymbol{r} | \phi_{i,0} \rangle \cdot \langle \phi_{i,0} | \boldsymbol{r} | \phi_{i,0} \rangle = \mathrm{Tr}(A_B U),$$

where the elements of A_B are

$$(A_B)_{ij} = \langle \phi_{j,0} | \boldsymbol{r} | \phi_{i,0} \rangle \cdot \langle \phi_{i,0} | \boldsymbol{r} | \phi_{i,0} \rangle.$$

Similarly, the A matrices for Pipek-Mezey and Edmiston-Ruedenberg are

$$(A_{PM})_{ij} = \sum_{A=1}^{N_A} \langle \phi_{j,0} | P_A | \phi_{i,0} \rangle \langle \phi_{i,0} | P_A | \phi_{i,0} \rangle,$$

$$(A_{ER})_{ij} = \langle \phi_{j,0} \phi_{i,0} | \frac{1}{r} | \phi_{i,0} \phi_{i,0} \rangle.$$

Therefore, a primitive iterative algorithm looks like:

1. For the current set of orbitals (matrix C), calculate A;
2. Compute the optimal U based on *Lemma 5*;
3. Update $C = CU$, and if $U \approx I$, stop, otherwise go back to Step 1.

The algorithm can be accelerated by extrapolation techniques, for example DIIS, as studied in Subotnik et al.'s paper[20]. It is worth mentioning that although *Lemma 5* promises that each line-search step we will move to the global maximum of the surrogate function, the algorithm does not promise to converge to the global maximum point of the problem. The computational cost of the algorithm for all three localization techniques scales as $O(N^3)$, where N is the number of the basis set functions.

As the determinant of A is not necessarily positive here (cf. the previous section where $\det A$ is always positive), we should briefly discuss this question. Since the occupied orbitals are orthogonal to each other, one can expect that A_{PM} and A_{ER} are strictly diagonal dominant. So, they must have positive determinants in most chemical systems because the diagonal elements of these two matrices are always positive. On the other hand, A_B, being heavily dependent on the position of the origin, does not share this property and may have a negative determinant. Figure 1 plots the smallest singular value (σ_m) of the relevant A matrices in some linear alkane systems. By the lemma, $2\sigma_m$ is the difference of the maximum value of $\mathrm{Tr}(AU)$ in $SO(n)$ and $O(n) - SO(n)$, a parity-violating difference. And, from the figure, it can be seen that restricting U to be on $SO(n)$ (which is what Subotnik et al. had conjectured) will not be inaccurate because either $\det A$ is always positive or σ_m is too small (around 10^{-6}). σ_m is also showing a monotonic decay as n increases for the cases of Pipek-Mezey and Edmiston-Ruedenberg. Finally, it needs to be noted that f_B, f_{PM}, f_{ER} are all even functions; in other words,

$$f\big(SO(n)\big) = f\big(O(n)\big).$$

It is only because we linearize f to f' that a parity-violating difference occurs.

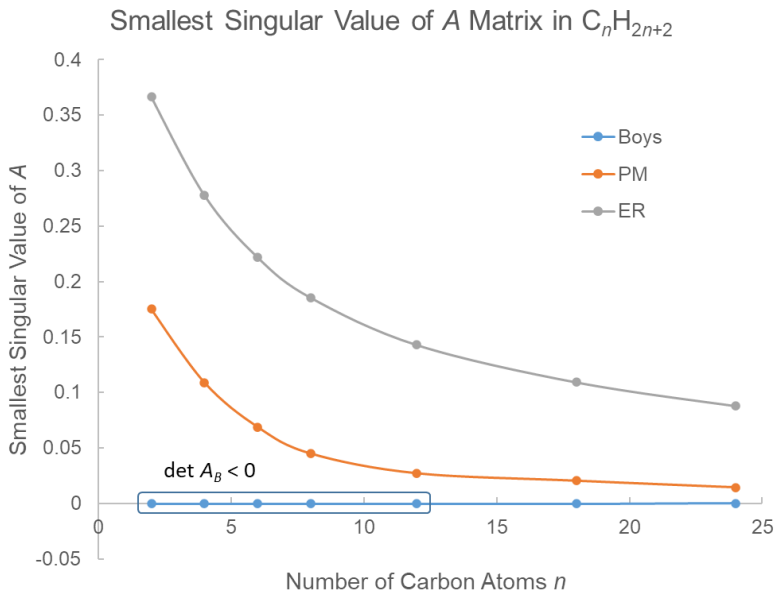

Figure 1 The smallest singular value of *A* matrix in some linear alkane systems. det *A* is always positive except for the boxed part of the Figure. PM stands for Pipek-Mezey and ER for Edmiston-Ruedenberg. The basis set is def2-SV(P).

3.3. Surrogate Optimization in Generalized Valence Bond Theory

The idea of preparing a linearized surrogate function in the previous section can also be applied to other branches of quantum chemistry. We will briefly look at an example of the active-active mixing in the minimal basis generalized valence bond (GVB) theory[39]. In contrast to conventional SCF orbital optimization, where only occupied-virtual orbital mixings affect the energy (Grassman manifold), all active-active rotations affect the GVB energy (Stieffel manifold) which typically makes optimization more challenging[18,40,41]. For simplicity, we will study a closed-shell system with frozen cores. The spatial part of the perfect pairing (PP) VB wavefunction is

$$\Psi = \boldsymbol{A}\left(\prod_{i=1}^{n_p}\left(c_{gi}\phi_{gi}\phi_{gi} + c_{ui}\phi_{ui}\phi_{ui}\right)\right),$$

where $\boldsymbol{A}$ is the antisymmetrizer, n_p is the number of electron pairs, $c_{gi}, c_{ui} > 0$ and $c_{gi}^2 + c_{ui}^2 = 1$. All ϕ orbitals are normalized and are orthogonal to each other. Goddard et al.[17,42] have derived the energy associated with this GVB-PP wavefunction, Ψ:

$$E = \sum_{i=1}^{n} 2f_i h_{ii}^{'} + \sum_{i,j=1}^{n} \left(a_{ij} J_{ij} + b_{ij} K_{ij}\right),$$

here $n = 2n_p$ is the number of active orbitals, $f_i = c_i^2$ is the orbital occupation coefficients, $h_{ii}^{'} = \langle \phi_i | \boldsymbol{h}^{'} | \phi_i \rangle$ and

$$\boldsymbol{h}^{'} = \boldsymbol{h} + \sum_{j=1}^{n_f} \left(2\boldsymbol{J}^j - \boldsymbol{K}^j\right).$$

In the last formula, n_f is the number of frozen (doubly occupied) orbitals, $\boldsymbol{h}$ is the usual one-electron Hamiltonian, and $\boldsymbol{J}^j, \boldsymbol{K}^j$ are standard Coulomb and exchange operators for frozen orbital j; $J_{ij} = \langle \phi_i \phi_j | \phi_i \phi_j \rangle, K_{ij} = \langle \phi_i \phi_i | \phi_j \phi_j \rangle$. The two-electron coefficients a_{ij}, b_{ij} are

$$a_{ij} = 2f_i f_j, b_{ij} = -f_i f_j, \quad \text{if } i, j \text{ are not in the same pair;}$$

$$a_{ij} = f_i, b_{ij} = 0, \quad \text{if } i = j;$$

$$a_{ij} = 0, b_{ij} = -\sqrt{f_i f_j}, \quad \text{if } i \neq j \text{ but are in the same pair.}$$

One should notice that E is a function of $\{\phi_i\}_{i=1}^n$ and $\{f_i\}_{i=1}^n$. As *Lemma 5* only finds the maximum of a function of an orthogonal matrix, using the lemma to simultaneously optimize $\{\phi_i\}_{i=1}^n$ and $\{f_i\}_{i=1}^n$ is not possible. We can, however, propose a two-step algorithm as below:

1. Fixing $\{f_i\}_{i=1}^n$, optimize $\{\phi_i\}_{i=1}^n$, which can be achieved by a surrogate function as will be described below;
2. Fixing $\{\phi_i\}_{i=1}^n$, optimize $\{f_i\}_{i=1}^n$, which is an algebraic, n_p variables, constrained optimization problem;
3. Check if self-consistency is reached, and if yes, stop, otherwise go back to Step 1.

We will not discuss Step 2 in detail as it is not the main topic of this paper.

For Step 1, we can linearize the energy function such that

$$E^{'}(U) = \sum_{i=1}^{n} f_i \langle \phi_i | \boldsymbol{h}^{'} | \phi_{i,0} \rangle + \sum_{i,j=1}^{n} \left(a_{ij} \langle \phi_i \phi_{j,0} | \phi_{i,0} \phi_{j,0} \rangle + b_{ij} \langle \phi_i \phi_{i,0} | \phi_{j,0} \phi_{j,0} \rangle\right),$$

where $\{\phi_{i,0}\}_{i=1}^{n}$ is a fixed set of orbitals. We eliminate the factor of 2 to compensate for the fact that the two terms in $E(U)$ have different order with respect to U (quadratic and quartic, respectively). Also,

$$\phi_i = \sum_{j=1}^{n} U_{ji}\phi_{j,0},$$

so

$$E'(U) = \sum_{i=1}^{n} f_i \sum_{k=1}^{n} U_{ki} \langle \phi_{k,0}|\boldsymbol{h}'|\phi_{i,0}\rangle + \sum_{i,j=1}^{n}\sum_{k=1}^{n} U_{ki}\big(a_{ij}\langle\phi_{k,0}\phi_{j,0}|\phi_{i,0}\phi_{j,0}\rangle + b_{ij}\langle\phi_{k,0}\phi_{i,0}|\phi_{j,0}\phi_{j,0}\rangle\big),$$

and by defining

$$H'_{ik} = f_i\langle\phi_{k,0}|\boldsymbol{h}'|\phi_{i,0}\rangle,$$

and

$$G_{ik} = \sum_{j=1}^{n}\big(a_{ij}\langle\phi_{k,0}\phi_{j,0}|\phi_{i,0}\phi_{j,0}\rangle + b_{ij}\langle\phi_{k,0}\phi_{i,0}|\phi_{j,0}\phi_{j,0}\rangle\big),$$

$E'(U)$ can be cast into a familiar form

$$E'(U) = \mathrm{Tr}(H'U) + \mathrm{Tr}(GU) = \mathrm{Tr}[(H' + G)U].$$

An iterative algorithm based on *Lemma 5* for Step 1 is therefore natural. When the algorithm converges, we have

$$(H' + G)^T\left((H' + G)(H' + G)^T\right)^{-1/2} = I,$$

meaning

$$(H' + G)^T = H' + G,$$

which coincides with the stationary point condition of the Fock operator; i.e.,

$$\langle\phi_i|\boldsymbol{F}^i - \boldsymbol{F}^j|\phi_j\rangle = 0, \quad \forall i, j,$$

where

$$\boldsymbol{F}^i = f_i\boldsymbol{h}' + \sum_{j=1}^{n}\big(a_{ij}\boldsymbol{J}^j + b_{ij}\boldsymbol{K}^j\big)$$

is the Fock operator of orbital i. The scaling of the algorithm is $O(n^3N)$, where N is the number of atomic orbital basis functions, as we need to do an integral transformation to build G.

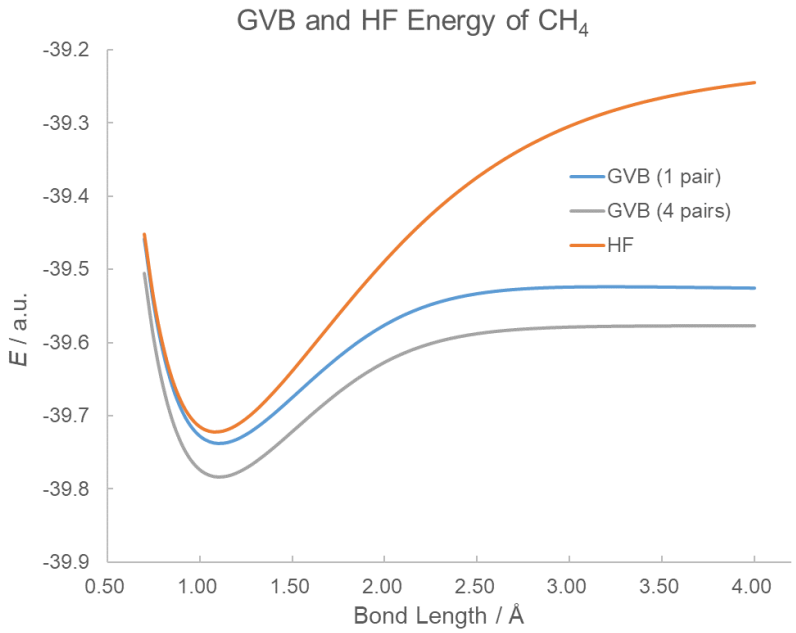

Figure 2 Restricted HF and GVB-PP energy of CH_4 with one of the C-H bonds stretched. "1 pair" and "4 pairs" stand for the size of the active space. The basis set is STO-3G.

We have tested the algorithm on CH_4 with the minimal STO-3G basis set (Figure 2). To achieve stable convergence, on every iteration in Step 2, instead of transforming by the optimal U, we use $U^{1/2}$, and we update $\{f_i\}_{i=1}^n$ after every 5 iterations of orbital optimization. The $\{f_i\}_{i=1}^n$ is optimized by the Rosen Gradient Projection method with Golden Section method on the line search steps. From the figure, it is clear that the 1-pair GVB treatment is enough to achieve a qualitatively correct potential energy surface, although of course the 4-pair GVB wavefunction indeed acquires more correlation energy than the 1-pair wavefunction. At the bond length of 2.00 Å, it takes about 1500 iterations to converge, which is much slower comparing with about 25 iterations when using the geometric direct minimization method (GDM)[18] – how to stably accelerate the algorithm is a direction of future study. It is worth pointing out that this algorithm can, theoretically, also be applied to HF calculation, where $\{f_i\}_{i=1}^n$ is fixed and we only need to optimize the

orbitals. However, this is not likely to be competitive with the Roothaan step, which exactly extremizes a surrogate function that is quadratic in U.

4. Conclusion

In this paper, we have presented the correct form and the complete proof of a lemma (*Lemma 5*), establishing that the function $f(U) = \mathrm{Tr}(AU)$ has one local maximum point each in $SO(n)$ and $O(n) - SO(n)$ with different function values. Various applications of the lemma in quantum chemistry were given, including orbital orthogonalization (Carlson-Keller theorem and its weighted counterpart), orbital localization (Boys, Pipek-Mezey, and Edmiston-Ruedenberg), and orbital optimization (generalized valence bond). The surrogate optimization technique associated with this lemma is proved to be effective in those cases. Although the lemma itself only gives the precise optimal point to a very specific type of surrogate function, and surrogate optimization in SCF is routine, we believe that the full power of surrogate optimization has not yet been utilized in quantum chemistry. There is scope for future research utilizing the lemma in different chemical problems, as well as proposing other surrogate functions and studying their properties.

Acknowledgements

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing, and Office of Basic Energy Sciences, via the Scientific Discovery through Advanced Computing (SciDAC) program. Z. W. thanks Haoyu Wang (Tsinghua University) for fruitful discussions on the proof.

Conflict of Interest

None.

References

[1] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.

[2] J. Behler, *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930.

[3] J. Tiihonen, P. R. C. Kent, J. T. Krogel, *J. Chem. Phys.* **2022**, *156*, 054104.

[4] G. Raggi, I. F. Galván, C. L. Ritterhoff, M. Vacher, R. Lindh, *J. Chem. Theory Comput.* **2020**, *16*, 3989–4001.

[5] B. Lei, T. Q. Kirk, A. Bhattacharya, D. Pati, X. Qian, R. Arroyave, B. K. Mallick, *npj Comput. Mater.* **2021**, *7*, 1–12.

[6] J. Leguy, B. Duval, B. Da Mota, T. Cauchy, *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI* **2021**, *2021-Novem*, 780–785.

[7] K. T. Schütt, M. Gastegger, A. Tkatchenko, K. R. Müller, R. J. Maurer, *Nat. Commun.* **2019**, *10*, 1–10.

[8] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, R. Ramprasad, *npj Comput. Mater.* **2019**, *5*, 22.

[9] P. Echenique, J. L. Alonso, *Mol. Phys.* **2007**, *105*, 3057–3098.

[10] A. T. B. Gilbert, N. A. Besley, P. M. W. Gill, *J. Phys. Chem. A* **2008**, *112*, 13164–13171.

[11] D. Hait, M. Head-Gordon, *J. Phys. Chem. Lett.* **2020**, *11*, 775–786.

[12] D. Hait, M. Head-Gordon, *J. Phys. Chem. Lett.* **2021**, *12*, 4517–4529.

[13] H. Werner, P. J. Knowles, *J. Chem. Phys.* **1985**, *82*, 5053–5063.

[14] D. L. Yeager, D. Lynch, J. Nichols, P. Joergensen, J. Olsen, *J. Phys. Chem.* **1982**, *86*, 2140–2153.

[15] D. A. Kreplin, P. J. Knowles, H. J. Werner, *J. Chem. Phys.* **2019**, *150*, 194106.

[16] D. A. Kreplin, P. J. Knowles, H. J. Werner, *J. Chem. Phys.* **2020**, *152*, 074102.

[17] L. G. Yaffe, W. A. Goddard, *Phys. Rev. A* **1976**, *13*, 1682–1691.

[18] S. Lehtola, J. Parkhill, M. Head-Gordon, *Mol. Phys.* **2018**, *116*, 547–560.

[19] W.-K. Tung, *Group Theory in Physics*, WORLD SCIENTIFIC, **1985**.

[20] J. E. Subotnik, Y. Shao, W. Z. Liang, M. Head-Gordon, *J. Chem. Phys.* **2004**, *121*, 9220–9229.

[21] B. C. Carlson, J. M. Keller, *Phys. Rev.* **1957**, *105*, 102–103.

[22] J. C. Gower, G. B. Dijksterhuis, *Procrustes Problems*, Oxford University Press, **2004**.

[23] P. H. Schönemann, *Psychometrika* **1966**, *31*, 1–10.

[24] W. Kabsch, *Acta Crystallogr. Sect. A* **1976**, *32*, 922–923.

[25] W. Kabsch, *Acta Crystallogr. Sect. A* **1978**, *34*, 827–828.

[26] S. Umeyama, *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*, 376–380.

[27] G. R. Kneller, *Mol. Simul.* **1991**, *7*, 113–119.

[28] G. Wahba, *SIAM Rev.* **1965**, *7*, 409–409.

[29] P. Löwdin, *J. Chem. Phys.* **1950**, *18*, 365–375.

[30] J. A. Goldstein, M. Levy, *Am. Math. Mon.* **1991**, *98*, 710–718.

[31] J. G. Aiken, J. A. Erdos, J. A. Goldstein, *Int. J. Quantum Chem.* **1980**, *18*, 1101–1108.

[32] I. Mayer, *Int. J. Quantum Chem.* **2002**, *90*, 63–65.

[33] R. Bhatia, K. K. Mukherjea, *Int. J. Quantum Chem.* **1986**, *29*, 1775–1778.

[34] D. F. Scofield, *Int. J. Quantum Chem.* **1973**, *7*, 561–568.

[35] S. F. Boys, *Rev. Mod. Phys.* **1960**, *32*, 296–299.

[36] J. M. Foster, S. F. Boys, *Rev. Mod. Phys.* **1960**, *32*, 300–302.

[37] J. Pipek, P. G. Mezey, *J. Chem. Phys.* **1989**, *90*, 4916–4926.

[38] C. Edmiston, K. Ruedenberg, *Rev. Mod. Phys.* **1963**, *35*, 457–464.

[39] W. A. Goddard, T. H. Dunning, W. J. Hunt, P. J. Hay, *Acc. Chem. Res.* **1973**, *6*, 368–376.

[40] A. Edelman, T. A. Arias, S. T. Smith, *SIAM J. Matrix Anal. Appl.* **1998**, *20*, 303–353.

[41] T. van Voorhis, M. Head-Gordon, *Mol. Phys.* **2002**, *100*, 1713–1721.

[42] R. P. Muller, J. M. Langlois, M. N. Ringnalda, R. A. Friesner, W. A. Goddard, *J. Chem. Phys.* **1994**, *100*, 1226–1235.