Global Mapping of Surface Composition on an Exo-Earth Using Sparse Modeling

Atsuki Kuwata1, Hajime Kawahara2,3, Masataka Aizawa4, Takayuki Kotani5,6,7, and Motohide Tamura1,5,6

1 Department of Astronomy, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan; 2 Department of Earth and Planetary Science, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan; 3 Research Center for the Early Universe, School of Science, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan; 4 Tsing-Dao Lee Institute, Shanghai Jiao Tong University, 520 Shengrong Road, Shanghai 201210, People’s Republic of China; 5 National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan; 6 Astrobiology Center, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan; 7 Department of Astronomy, School of Science, The Graduate University for Advanced Studies, SOKENDAI, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan

Received 2022 February 15; revised 2022 March 24; accepted 2022 April 4; published 2022 May 17

Abstract

The time series of light reflected from exoplanets by future direct imaging can provide spatial information with respect to the planetary surface. We apply sparse modeling to the retrieval method that disentangles the spatial and spectral information from multiband reflected light curves termed as spin–orbit unmixing. We use the ℓ1-norm and the total squared variation norm as regularization terms for the surface distribution. Applying our technique to a toy model of cloudless Earth, we show that our method can infer sparse and continuous surface distributions and also unmixed spectra without prior knowledge of the planet surface. We also apply the technique to the real Earth data as observed by DSCOVR/EPIC. We determined the representative components that can be interpreted as cloud and ocean. Additionally, we found two components that resembled the distribution of land. One of the components captures the Sahara Desert, and the other roughly corresponds to vegetation, although their spectra are still contaminated by clouds. Sparse modeling significantly improves the geographic retrieval, in particular, of clouds and leads to higher resolutions for other components when compared with spin–orbit unmixing using Tikhonov regularization.

Unified Astronomy Thesaurus concepts: Astrobiology (74); Astroinformatics (78); Exoplanet astronomy (486); Direct imaging (387); Exoplanet surfaces (2118); Exoplanet surface characteristics (496); Exoplanet surface composition (2022)

1. Introduction

The photometric variation of directly imaged exoplanets has been considered as an invaluable probe for the environment of habitable planets, as well as spectroscopy (Ford et al. 2001). A two-dimensional surface distribution can be decoded from the diurnal and seasonal variations in reflected light (Kawahara & Fujii 2010). To date, this technique, termed as spin–orbit tomography, has been studied in terms of the regularization of geography (Kawahara & Fujii 2011; Fujii & Kawahara 2012), Bayesian formulation (Farr et al. 2018; Kawahara & Masuda 2020), a planet’s axial tilt determination (Schwartz et al. 2016; Farr et al. 2018), dynamical mapping (Kawahara & Masuda 2020), non-Lambertian effect (Luger et al. 2021), and its application to real Earth data (Fan et al. 2019; Luger et al. 2019). Furthermore, the importance of regularization has been recognized. Tikhonov regularization, which was originally applied by Kawahara & Fujii (2011) to spin–orbit tomography, tends to exhibit a smooth solution as an inferred map. Aizawa et al. (2020) showed that sparse modeling improved the sharpness of the inferred map. Sparse modeling is an optimization technique that assumes the sparsity of the solution, as described in Appendix C. More recently, Asensio Ramos & Pallé (2021) introduced a neural-net-based regulator, learned denoiser, which also improved map quality.

Spin–orbit tomography is formulated to retrieve a two-dimensional spatial map of a single component such as single-band photometry, color difference, and a single principle component. On the other hand, multicolor photometric variation contains information on the spectra of individual surface components such as water, soil, vegetation, snow, and clouds. Rotational spectral unmixing is a blind retrieval of end-member spectra of individual surface components combined with the disentanglement of geography by spin rotation, and it has been explored using EPoxy satellite data (Cowan & Strait 2013; Lustig-Yaeger et al. 2018). However, rotational spectral unmixing suffers from nonuniqueness of the inferred end-member spectra (Fujii et al. 2017). Recently, Kawahara (2020) proposed spin–orbit unmixing, which is a unified scheme of spin–orbit tomography and spectral unmixing that leverages a nonnegative matrix factorization (NMF) and simplex volume minimization. The latter technique is essentially regularization of the spectral components proposed in the remote sensing field (Craig 1994), which eliminates the ambiguity of the end-member determination. However, the regularization for geography is still a traditional Tikhonov (ℓ2-norm) regularization in the current spin–orbit unmixing. In this study, we show that sparse modeling in spin–orbit unmixing improves not only the map quality but also the recovery of the end-member spectra.

The remainder of this paper is organized as follows. We review the basic formulation of spin–orbit unmixing in Section 2. In Section 3, we formulate the spin–orbit unmixing with sparsity. Furthermore, in Section 4, using a toy model, we verify our technique. In Section 5, we demonstrate the new technique by applying it to real observational data of Earth obtained from the Deep Space Climate Observatory.
The planet flux of reflected light from a solid surface is expressed (see Appendix A in Kawahara 2020, for the derivation) as follows:
\[ f_p^{\text{ref}} = \frac{f_s R_p^2}{\pi a^2} \int_{S_1 \cap S_V} dS \theta(\theta_0, \varphi_0, \theta_1, \varphi_1)(e_S \cdot e_R)(e_O \cdot e_R). \]
(1)

where \( R_p \) denotes a planet radius; \( a \) is a star–planet distance; \( f_s \) is the stellar flux; \( R(\theta_0, \varphi_0, \theta_1, \varphi_1) \) is the bidirectional reflectance distribution function; \( \theta_0 \) and \( \varphi_0 \) are the incident zenith angle and azimuth angle, respectively; \( \theta_1 \) and \( \varphi_1 \) are the reflected zenith angle and azimuth angle, respectively; \( e_O, e_S, \) and \( e_R \) denote the unit vector from the center of the planet to the observer, primary star, and surface, respectively; and \( S_1 \cap S_V \) denotes the overlapped area of the illuminated region \( S_1 \) \((e_S \cdot e_R > 0)\) and the visible region \( S_V \) \((e_O \cdot e_R > 0)\). Assuming isotropic reflection (Lambertian), the local reflectivity on the planet surface is expressed as a function of the spherical coordinate:
\[ m(\theta, \phi) = R(\theta_0, \varphi_0, \theta_1, \varphi_1). \]
(2)

Then, we obtain
\[ f_p^{\text{ref}}(t) = \int dSG(t, \theta, \phi)m(\theta, \phi), \]
(3)

where
\[ G(t, \theta, \phi) = \begin{cases} \frac{f_s R_p^2}{\pi a^2}(e_S \cdot e_R)(e_O \cdot e_R) \\ (e_S \cdot e_R > 0, e_O \cdot e_R > 0) \\ 0 \quad \text{(otherwise).} \end{cases} \]
(4)

The discretization of \( t \) and \( (\theta, \phi) \) yields
\[ f_p^{\text{ref}}(t_i) = \sum_j \Delta SG(t_i, \theta_j, \phi_j)m(\theta_j, \phi_j). \]
(5)

In this study, we pixelized the planet surface using Hierarchical Equal Area isoLatitude Pixelization (HEALPix; Górski et al. 2005). Then, Equation (5) is written in matrix form as follows:
\[ d = Wm, \]
(6)

where \( d_i := f_p^{\text{ref}}(t_i); \) \( m_j := m(\theta_j, \phi_j); \) \( W_{ij} := G(t_i, \theta_j, \phi_j) \Delta S \)

\( (i = 1, \ldots, N_t, j = 1, \ldots, N_j); \) and \( N_t \) and \( N_j \) are the number of observation data and pixels of the surface, respectively. We infer the surface distribution of planet \( m_{\text{est}} \) from the observation data \( d_{\text{obs}} \). In reality, we must include an observation error \( \varepsilon \) in the model:
\[ d_{\text{obs}} = Wm + \varepsilon. \]
(7)

Spin–orbit tomography infers \( m \) from \( d_{\text{obs}} \). In general, \( W \) is a function of an orbital inclination \( i \), an orbital phase at equinox \( \Theta_{\text{eq}} \), and a planet spin vector or equivalently a set of planet obliquity \( \zeta \). Additionally, the planet rotation period should be inferred. These nonlinear parameters can be inferred within the framework of spin–orbit tomography (Kawahara & Fujii 2010; Schwartz et al. 2016; Farr et al. 2018; Kawahara & Masuda 2020). In this study, we assume that these nonlinear parameters are known. This type of optimization problem then becomes a linear inverse problem.

In general, the linear inverse problem requires regularization of the model to suppress the model instability due to noise, namely, to avoid overfitting or overlearning. Regularization can take a variety of formulations, such as a bounded model (Kawahara & Fujii 2010), a nonnegative condition (Kawahara 2020), a regularization term (Kawahara & Fujii 2011; Aizawa et al. 2020), and neural network (Asensio Ramos & Pallé 2021). Using the regularization term, the inverse problem can be formulated as an optimization problem:
\[ \min m \frac{1}{2} \|d - Wm\|^2_2 + \gamma R(m), \]
(8)

Kawahara & Fujii (2011) used Tikhonov regularization as follows:
\[ \min m \frac{1}{2} \|d - Wm\|^2_2 + \lambda \text{TSV}_\gamma \|m\|^2_2, \]
(9)

where \( \lambda \) denotes the regularization parameter. Tikhonov regularization tends to provide a smoother solution than the other regularization methods. To improve the sharpness of the inferred map, Aizawa et al. (2020) introduced a combination of \( \ell_1 \)-norm regularization and total squared variation (TSV) regularization (Kuramochi et al. 2018) as a sparse modeling as follows:
\[ \min m \frac{1}{2} \|d - Wm\|^2_2 + \lambda_0 \|m\|_1 + \lambda_{\text{TSV}} \|m\|_{\text{TSV}}, \]
(10)

where \( \lambda_0 \) and \( \lambda_{\text{TSV}} \) denote the regularization parameters and \( \| \cdot \|_1 \) and \( \| \cdot \|_{\text{TSV}} \) denote \( \ell_1 \)-norm and TSV norm, respectively.

Spin–orbit tomography can retrieve a map of a single component of a light curve. For instance, Kawahara & Fujii (2011) used the color difference between the near-infrared and visible bands to reduce the cloud contribution. The retrieved map accurately represented the land and ocean distribution. Fan et al. (2019) used a second principle component of the principal component analysis, which also recovered the land/ocean distribution from the real data of Earth obtained via DSCOVR.

To infer map and surface components, Cowan & Strait (2013) applied spectral unmixing to the longitudinal mapping of Earth. However, the retrieved map did not match the real land and ocean distribution. Kawahara (2020) extended their method to a two-dimensional mapping, corresponding to a combination of spectral unmixing and spin–orbit tomography, namely, spin–orbit unmixing. They also introduced regularization terms for geography and spectra. The details of spectral unmixing are presented in Appendix A.

### 2.2. Spin–Orbit Unmixing

We explain spin–orbit unmixing. Spin–orbit unmixing uses the observation data matrix \( D \in \mathbb{R}^{N_t \times N_j} \) along the time and wavelength axes. Let \( M \in \mathbb{R}^{N_t \times N_j} \) be the surface distribution defined as \( M_{ij} := m(\theta_j, \phi_j, \lambda_0) \), and let \( A \in \mathbb{R}^{N_t \times N_j} \) be the surface distribution matrix and end-member matrix, respectively. Then, the formulation of the spin–orbit
unmixing is as follows:

\[ D = W A X. \]  

(11)

As an optimization problem, the spin–orbit unmixing can be expressed as follows:

\[
\begin{align*}
\text{minimize } & \frac{1}{2} \| D - W A X \|_F^2 + R(A, X) \\
\text{subject to } & A \times k \geq 0, X \delta l \geq 0,
\end{align*}
\]

(12)

where \( R(A, X) \) denotes a regularization term and \( \| \cdot \|_F \) denotes the Frobenius norm, which can be defined as

\[
\| Y \|_F = \left( \sum_{i} \sum_{j} Y_{ij}^2 \right)^{1/2}.
\]

(13)

Kawahara (2020) used Tikhonov regularization for \( A \) and determinant type of the volume regularization for \( X \) as follows:

\[
\begin{align*}
\text{minimize } & \frac{1}{2} \| D - W A X \|_F^2 + \lambda_A \| A \|_F^2 + \lambda_X \det(XX^T) \\
\text{subject to } & A \times k \geq 0, X \delta l \geq 0,
\end{align*}
\]

(14)

where \( \lambda_A \) and \( \lambda_X \) denote the regularization parameters. Volume regularization is based on the concept of simplex volume minimization, which was developed in the field of remote sensing (Craig 1994; Fu et al. 2015, 2019; Lin et al. 2015; Ang & Gillis 2019). With respect to spectral unmixing, it is known that NMF with the volume regularization term accurately reproduces high-resolution spectrum components from satellite data. Simplex volume minimization is justified under the assumption that the data points are widely spread in the convex hull defined by the end-members. The true end-members are then identified by the data-enclosing simplex, whose volume is minimized (Craig 1994). An intuitive explanation is provided in Figure 1 in Lin et al. (2015) and Figure 1 in Kawahara (2020). The choice of \( \det(XX^T) \) as a regularization term is based on the following fact: the volume of an \((N_i - 1)\)-simplex (or convex hull) in \((N_i - 1)\)-dimensional space with vertices \( \{ x_1, x_2, \ldots, x_N \} \) is \( \det(XX^T)/(N_i!) \), where \( x_4 \) denotes the 4th end-member (i.e., the \( k \)th column vector) of \( X^T \).

3. Spin–Orbit Unmixing with Sparsity

In this study, we introduce the sparsity of geography \( A \) into spin–orbit unmixing. Most of the elements are zero in a sparse matrix. Sparse optimization is described in Appendix C. The objective function in this study is expressed as follows:

\[
\begin{align*}
\text{minimize } & Q \text{ subject to } A \times k \geq 0, X \delta l \geq 0, \\
Q := & \frac{1}{2} \| D - W A X \|_F^2 + R(A, X).
\end{align*}
\]

(15)

As a regularization of \( A \), we consider two types of sparse modeling: \( \ell_1 \)-norm+TSV, and trace norm regularization. We also computed Tikhonov regularization for comparison purposes.

In Equation (15), we must optimize \( A \) and \( X \). Based on Kawahara (2020), we used the block coordinate descent method, which divides the problem into two separate optimizations for \( A \) and \( X \) (e.g., Kim et al. 2014).

\[
\begin{align*}
\text{minimize } & q_A \text{ subject to } A \times k \geq 0, \\
q_A := & \frac{1}{2} \| P - W A k \|_F^2 + \lambda_T \| A \|_F^2 + \lambda_X \det(XX^T).
\end{align*}
\]

(16)

where \( q_A = q_A(A) \) and \( q_X = q_X(X) \) are defined by rearranging Equation (16) as

\[
\begin{align*}
Q &= q_A(A) + (\text{constant for } A), \\
Q &= q_X(X) + (\text{constant for } X),
\end{align*}
\]

(17)

and these optimizations are iteratively conducted until convergence.

### 3.1. Spin–Orbit Unmixing with \( \ell_1 \)-norm and TSV Regularization

In the context of spin–orbit tomography, Aizawa et al. (2020) described a procedure for obtaining a sparse solution using the \( \ell_1 \)-norm and TSV regularization. We extend the optimization method to spin–orbit unmixing. The optimization problem with the \( \ell_1 \)-norm and TSV regularization can be expressed as follows:

\[
\begin{align*}
\text{minimize } & Q_{\ell + \text{TSV}} \text{ subject to } A \times k \geq 0, \\
Q_{\ell + \text{TSV}} := & \frac{1}{2} \| D - W A X \|_F^2 \\
& + \sum_k (\lambda_{\ell} \| a_k \|_1 + \lambda_{\text{TSV}} \| a_k \|_{\text{TSV}}) \\
& + \lambda_X \det(XX^T),
\end{align*}
\]

(18)

where \( a_k \) denotes the \( k \)th column vector of \( A \) and \( \lambda_{\ell}, \lambda_{\text{TSV}}, \) and \( \lambda_X \) denote the regularization parameters. We used the determinant type of volume regularization for \( X \) as introduced in Section 2.2.

First, we solve optimization problem (21) for \( A \). \( Q_{\ell + \text{TSV}} \) is rearranged with respect to \( a_k \) as follows:

\[
\begin{align*}
Q_{\ell + \text{TSV}} &= \frac{1}{2} \| x_k \|_2^2 \| p_l - W a_k \|_2^2 + \lambda_{\ell} \| a_k \|_1 \\
&+ \lambda_{\text{TSV}} \| a_k \|_{\text{TSV}} + (\text{constant for } a_k),
\end{align*}
\]

(19)

where \( x_k \) is the \( k \)th column vector of \( X^T \), \( p_l := (1/\| x_k \|_2^2) \Delta x_k \), and \( \Delta \) is a matrix defined as \( \Delta_{ij} := D_{ij} - \sum_{j=1}^{N} W_{ij} A_{ij} X_{ij} \). Therefore, the subproblem for solving optimization problem (21) is expressed as follows:

\[
\begin{align*}
\text{minimize } & q_{A,\ell + \text{TSV}} \text{ subject to } (a_k)_j \geq 0 (j = 1, \ldots, N), \\
q_{A,\ell + \text{TSV}} := & \frac{1}{2} \| P - W a_k \|_F^2 \\
&+ \lambda_{\ell} \| a_k \|_1 + \lambda_{\text{TSV}} \| a_k \|_{\text{TSV}},
\end{align*}
\]

(20)

where \( \lambda_{\ell} := \lambda_{\ell}/\| x_k \|_2^2 \) and \( \lambda_{\text{TSV}} := \lambda_{\text{TSV}}/\| x_k \|_2^2 \). Then, we suppose

\[
\begin{align*}
f_{\ell + \text{TSV}}(a_k) := & \frac{1}{2} \| P - W a_k \|_F^2 + \lambda_{\ell} \| a_k \|_1 + \lambda_{\text{TSV}} \| a_k \|_{\text{TSV}}, \\
\psi_{\ell + \text{TSV}}(a_k) := & \lambda_{\ell} \| a_k \|_1 + \delta_{\epsilon}(a_k),
\end{align*}
\]

(21)

where \( \delta_{\epsilon} \) is defined in Equation (B19). It should be noted that \( f_{\ell + \text{TSV}}(a_k) \) is differentiable and \( \psi_{\ell + \text{TSV}}(a_k) \) is nondifferentiable. By adopting the above equation, Equation (24) is
rewritten as follows:

\[
\min_{a_k} f_{\ell^1+TSV}(a_k) + \psi_{\ell^1+TSV}(a_k). \tag{28}
\]

This objective function consists of a combination of two proper convex functions (Definition 3 in Appendix F) that are differentiable and not necessarily differentiable. This type of objective function can be optimized using the proximal gradient method, as explained in Appendix B.2.

The gradient of \( f_{\ell^1+TSV}(a_k) \) and proximal operator of \( \psi_{\ell^1+TSV}(a_k) \) can be obtained as follows:

\[
\nabla f_{\ell^1+TSV}(a_k) = -W^T(p_a - Wa_k) + 2\lambda_{\ell^1+TSV}\tilde{N}a_k, \tag{29}
\]

\[
\text{prox}(w|\psi_{\ell^1+TSV}) = \max \{ w - \gamma\lambda_{\ell^1}, 0 \}. \tag{30}
\]

where \( \tilde{N} \) is defined in Equation (C11), \( \gamma > 0 \) is the parameter indicating the step size, \( \mathbf{1} = (1, 1, \ldots, 1)^T \) denotes a vector such that all entries are one (one vector), and \( \max \) denotes the element-wise maximum. Hence, the update formula of the proximal gradient method for Equation (24) can be expressed as follows:

\[
a^{(i+1)}_k = \max \{ a^{(i)}_k - \gamma x^{(i)}_k - \gamma \lambda_{\ell^1}, 0 \}, \tag{31}
\]

\[
s^{(i)}_k = -W^T(p_a - Wa^{(i)}_k) + 2\lambda_{\ell^1+TSV}\tilde{N}a^{(i)}_k. \tag{32}
\]

In this study, we used Monotone FISTA (MFISTA; Beck & Teboulle 2009a) to solve the subproblem for \( a_k \) (Equation (24)). MFISTA deals with the nonmonotonic decreasing nature of the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA; Beck & Teboulle 2009b), and Aizawa et al. (2020) used it to solve spin–orbit tomography with the \( \ell_1 \)-norm and TSV regularization.

Next, we consider optimization problem (21) for \( X \). The objective function \( Q_{\ell^1+TSV} \) can be rearranged with respect to \( x_k \) as follows:

\[
Q_{\ell^1+TSV} = \frac{1}{2} x_k^T(L_X + D_X)x_k - \ell^T_k x_k + \text{(constant for } x_k), \tag{33}
\]

where \( L_X = ||W_a||_F^2 I, D_X = \lambda_X \det(\tilde{x}_k\tilde{x}_k^T)(I - \tilde{x}_k(\tilde{x}_k\tilde{x}_k^T)^{-1}\tilde{x}_k) \), \( \tilde{x}_k = \Delta^T W_a, I \) is an identity matrix, and \( \tilde{x}_k \) is a submatrix of \( X \) with the 4th row removed. The subproblem for solving optimization problem (21) is expressed as follows:

\[
\text{minimize } q_{X,\ell^1+TSV} \quad \text{subject to } (x_k)_l \geq 0 \quad (l = 1, \ldots, N), \tag{34}
\]

\[
q_{X,\ell^1+TSV} := \frac{1}{2} x_k^T(L_X + D_X)x_k - \ell^T_k x_k. \tag{35}
\]

Subproblem (34) is an optimization problem with a non-negative constraint for a differentiable objective function. This problem can be optimized using the proximal gradient method as explained in Appendix B.2. The derivative with respect to \( x_k \) can be obtained as follows:

\[
\nabla q_{X,\ell^1+TSV}(x_k) = (L_X + D_X)x_k - I_X. \tag{36}
\]

Then, the update formula of the proximal gradient method for (34) is written as follows:

\[
x_k^{(i+1)} = \max \{ x_k^{(i)} - \gamma ((L_X + D_X)x_k^{(i)} - I_X), 0 \}. \tag{37}
\]

Kawahara (2020) optimized subproblem (34) using FISTA, which is a proximal gradient method using Nesterov’s acceleration method (Nesterov 2003) with the restart method. Nesterov’s acceleration method increased the convergence speed. However, the function value did not necessarily decrease monotonically. The restart method (O’donoghue & Candés 2015) avoids increasing the objective function by restarting Nesterov’s acceleration method when the function value does not decrease.

Consequently, the optimization algorithm for solving (21) is summarized as follows:

### Algorithm 1. Spin–orbit unmixing with \( \ell_1 \)-norm and TSV regularization

| Initialization: nonnegative matrices \( A_0 \) and \( X_0 \) |
| for \( n \) in \( (1, N_n) \) do |
| for \( k \) in \( (1, N_k) \) do |
| Update \( x_k \) using FISTA with the restart method |
| Update \( a_k \) using MFISTA |
| end for |
| end for |

#### 3.2. Spin–Orbit Unmixing with Trace Norm Regularization

We also consider another type of sparse modeling, termed as trace norm regularization. Trace norm regularization is sparse modeling of a matrix, not a vector. It results in a low-rank matrix. Therefore, vectorization of a matrix, not a vector. It results in a low-rank matrix. Then, the update formula of the proximal gradient method for (34) is written as follows:

\[
x_k^{(i+1)} = \max \{ x_k^{(i)} - \gamma ((L_X + D_X)x_k^{(i)} - I_X), 0 \}. \tag{37}
\]

Kawahara (2020) optimized subproblem (34) using FISTA, which is a proximal gradient method using Nesterov’s acceleration method (Nesterov 2003) with the restart method. Nesterov’s acceleration method increased the convergence speed. However, the function value did not necessarily decrease monotonically. The restart method (O’donoghue & Candés 2015) avoids increasing the objective function by restarting Nesterov’s acceleration method when the function value does not decrease.

Consequently, the optimization algorithm for solving (21) is summarized as follows:

### Algorithm 1. Spin–orbit unmixing with \( \ell_1 \)-norm and TSV regularization

| Initialization: nonnegative matrices \( A_0 \) and \( X_0 \) |
| for \( n \) in \( (1, N_n) \) do |
| for \( k \) in \( (1, N_k) \) do |
| Update \( x_k \) using FISTA with the restart method |
| Update \( a_k \) using MFISTA |
| end for |
| end for |

#### 3.2. Spin–Orbit Unmixing with Trace Norm Regularization

We also consider another type of sparse modeling, termed as trace norm regularization. Trace norm regularization is sparse modeling of a matrix, not a vector. It results in a low-rank matrix. Therefore, vectorization of a matrix, not a vector. It results in a low-rank matrix. Then, the update formula of the proximal gradient method for (34) is written as follows:

\[
x_k^{(i+1)} = \max \{ x_k^{(i)} - \gamma ((L_X + D_X)x_k^{(i)} - I_X), 0 \}. \tag{37}
\]
in Appendix B.2. The derivative of $f_{\text{Tr}}(A)$ and the proximal operator of $\psi_{\text{Tr}}(A)$ are as follows:

$$
\nabla f_{\text{Tr}}(A) = -W^T(D - WA\lambda X)^T,
$$

(42)

$$
\text{prox}(W|\gamma \psi_{\text{Tr}}) = U \max \{\Sigma - \gamma \lambda_A I, O\} V^T,
$$

(43)

where $W = U\Sigma V^T$ denotes the singular value decomposition of $W$ (Lemma 8.4 in Tomioka 2015). Hence, the update formula of the proximal gradient method for (40) can be expressed as follows:

$$
A_{i+1} = U_i \max \{\Sigma_i - \gamma \lambda_A I, O\} V_i^T,
$$

(44)

where $Y_i = U_i \Sigma_i V_i^T$ denotes the singular value decomposition of $Y_i = A_i - \gamma \nabla f_{\text{Tr}}(A_i) = A_i + \gamma W^T(D - WA_i X)^T$. Hence, the optimization algorithm for solving (39) can be summarized as follows:

**Algorithm 2. Spin–orbit unmixing with trace norm regularization**

Initialization: nonnegative matrices $A_0$ and $X_0$

for $n$ in $(1, N_l)$ do

  for $k$ in $(1, N_k)$ do

    Update $X_k$ using FISTA with the restart method

  end for

  Update $A$ using FISTA with the restart method

end for

The code for optimization is publicly available.8

4. Testing Using a Cloudless Model

In this section, we describe the testing of our method using a cloudless Earth model. To compare with the results obtained using sparse modeling, we also consider spin–orbit unmixing with Tikhonov regularization and the determinant type of volume regularization (Kawahara 2020). The optimization algorithm for the latter is provided in Appendix D.

4.1. Cloudless Earth

First, we generate mock multicolor light curves as follows:

$$
D = W_{\text{true}} X_{\text{true}} + E,
$$

(45)

where $A_{\text{true}} \in \mathbb{R}^{N_f_{\text{true}} \times N_k_{\text{true}}}$ and $X_{\text{true}} \in \mathbb{R}^{N_k_{\text{true}} \times N}$ denote the input geography and surface spectra, respectively. The time interval is 1 yr divided by $N_i = 512$. Furthermore, $N_{f_{\text{true}}} = 12,288$, $N_{k_{\text{true}}} = 3$, and $N_k = 10$ denote the number of pixels on the planet surface, surface components, and observing bands, respectively. We considered vegetation, land, and ocean as the end-members of the surface components. Error matrix $E \in \mathbb{R}^{N_k \times N}$ is randomly generated as follows:

$$
E_{ij} \sim 0.01 D_{\text{true}} N(0, 1),
$$

(46)

$$
D_{\text{true}} := \frac{1}{N_k N_i} \sum_j (W_{\text{true}} X_{\text{true}})_{ij},
$$

(47)

where $N(0, 1)$ is a normal distribution with a mean $\mu = 0$ and standard deviation $\sigma = 1$. We used the classification map provided by the Moderate Resolution Imaging Spectroradiometer (MODIS) as the input surface distribution $A_{\text{true}}$, gray, white, and black correspond to vegetation, land, and ocean, respectively. (b) Input spectrum $X_{\text{true}}$.

Spectroradiometer (MODIS) as the input surface distribution $A_{\text{true}}$, the ASTER spectral library (Baldridge et al. 2009) for the spectra of vegetation and land, and the spectra produced by McLinden et al. (1997) for the ocean. We set the observation wavelength to $0.425 + 0.05(i - 1) \mu m (i = 1, \ldots, N_i = 10)$. $A_{\text{true}}$ and $X_{\text{true}}$ are shown in Figure 1. We also generated $W$ using the method described in Section 2.1, with orbital inclination $i = 45^\circ$, orbital phase angle at the vernal equinox $\Theta_{eq} = 90^\circ$, obliquity $\zeta = 23^\circ$, orbital period $P_{\text{orb}} = 365$ days, and rotation period $P_{\text{spin}} = 23.9344699/24.0$ days.

Based on $D$ and $W$, we infer $A \in \mathbb{R}^{N_f \times N_k}$ and $X \in \mathbb{R}^{N_k \times N}$ by solving the following optimization problem:

$$
\begin{array}{ll}
\text{minimize} & \frac{1}{2} \|D - WAX\|^2 + R(A, X) \\
\text{subject to} & A_{jk} \geq 0, X_{ij} \geq 0,
\end{array}
$$

(48)

where $R(A, X)$ takes the form in Equation (22) ($l_1$-norm and TSV regularization), Equation (39) (trace norm regularization), or Equation (14) (Tikhonov regularization). We set the number of pixels in the inferred map to $N_f = 3024$ and the number of end-members to $N_k = 3$.

4.2. Results of Spin–Orbit Unmixing with $l_1$-norm and TSV Regularization

We solved Equation (22) with $\lambda_{l_1} = 10^{-3.5}$, $\lambda_{TSV} = 10^{-4}$, $\lambda_X = 10^2$, and iteration number $N_{\text{try}} = 10^3$. These parameters

---

8 https://github.com/atsuki-kuwata/exomap
were selected based on the mean removed spectral angle (MRSA) and the correct pixel rate (CPR) as in Kawahara (2020). The detailed procedure is presented in Appendix E.

There is a known indefiniteness of matrix factorization, as explained below. The surface distribution at each wavelength $M \in \mathbb{R}^{N \times N}$ can be written as

$$M = AX = \sum_{k=1}^{N} a_k x_k,$$

(49)

where $a_k$ and $x_k$ ($k = 1, \ldots, N_k$) denote the $k$th column vectors of $A$ and $X^T$, respectively. By using the constant, $c_k$, we obtain the following:

$$a_k x_k = (c_k^{-1} a_k)(c_k x_k), (k = 1, \ldots, N_k).$$

(50)

This implies that the inferred surface distribution and spectrum have an indefiniteness of constant multiples. Hence, we normalize the inferred surface distribution and spectrum for $X$ as follows:

$$\hat{a}_k = \left( \frac{x_k^\text{true}}{x_k^T} \right)^{-1} a_k,$$

(51)

$$\hat{x}_k = \left( \frac{x_k^\text{true}}{x_k^T} \right) x_k,$$

(52)

where $x_k^\text{true}$ ($k = 1, \ldots, N_k^\text{true} = N_k$) is the $k$th column vector of $X^\text{true}$. Furthermore, $\bar{x}_k$ and $\bar{x}_k^\text{true}$ denote the means of $x_k$ and $x_k^\text{true}$, respectively:

$$\bar{x}_k = \frac{1}{N_k} \sum_{j=1}^{N_k} (x_k)_j,$$

(53)

$$\bar{x}_k^\text{true} = \frac{1}{N_k^\text{true}} \sum_{j=1}^{N_k^\text{true}} (x_k^\text{true})_j.$$  

(54)

The normalized surface distribution $\hat{A}$ and spectra $\hat{X}$ are shown in Figure 2. The inferred map accurately reproduces the structure of the input. Notably, the inferred spectra are in excellent agreement with the inputs, which is much better than the case of spin–orbit unmixing with Tikhonov regularization (Kawahara 2020).

4.3. Results of Spin–Orbit Unmixing with Trace Norm Regularization

Next, we tested spin–orbit unmixing with trace norm regularization by solving Equation (39) with $\lambda_A = 10^{3.1}$, $\lambda_X = 10^{3.3}$, and iteration number $N_{\text{try}} = 5 \times 10^2$.

We used the same evaluation measures as those in Appendix E to select the regularization parameters $\lambda_A$ and $\lambda_X$. Furthermore, we adopted $\lambda_A$ and $\lambda_X$ with a local minimum value of MRSA similar to that in Appendix E.

In the same manner as in Section 4.2, the normalized surface distribution $\hat{A}$ and spectra $\hat{X}$ are shown in Figure 3.

4.4. Comparison by Varying the Regularization Term

We compare the inferred solutions with $\ell_1$-norm and TSV regularization, trace norm regularization, and Tikhonov regularization (Figure 4). Figures 4(d)–(f) are the inferred surface distributions by spin–orbit unmixing with $\ell_1$-norm and TSV regularization. We can see less noise in the area with zero values and a more continuous surface with values in each end-member than the case with other regularizations. The sparseness and continuousness of the inferred map are induced by $\ell_1$-norm and TSV norm regularization, respectively. In

Figure 2. Surface map and spectra inferred by spin–orbit unmixing with the $\ell_1$-norm and TSV regularization using a cloudless Earth model. (a) Color composite of inferred surface distribution $\hat{A}$. (b) Inferred spectrum $\hat{X}$. Solid and dotted lines denote the inferred and input spectra, respectively. The inferred map for each end-member is displayed in (c) for vegetation, (d) for land, and (e) for the ocean.
general, it is difficult to retrieve the geography that contributes less to the data, such as vegetation in Oceania and land in South Africa, whichever regularization we use. However, the land distribution in Chile inferred with the $\ell_1$-norm and TSV regularization seems consistent with the input map, while the one with Tikhonov norm regularization is equivalent to the noise. Additionally, the inferred maps obtained using spin–orbit unmixing with $\ell_1$-norm and TSV regularization (Figures 4(d)–(f)) are smoother than those with Tikhonov regularization (Figures 4(j)–(l)). This could be because of the differences in properties between TSV and Tikhonov regularizations, both of which tend to exhibit smooth solutions. TSV regularization induces a smooth map by minimizing the square of the difference between the values of neighboring pixels (described in Appendix C.2), while Tikhonov regularization induces a smooth solution by preventing overfitting. Furthermore, the choice of regularization parameters and the nonnegative condition can affect the smoothness and noise of the inferred map. The results can be compared using various regularizations or constraints in future studies.

In the inference of the spectra (Figures 4(m)–(o)), we used simplex volume regularization in all cases. Although regularization terms for the spectra are the same, the inferred spectra are affected by the change in the regularization term for the surface distribution. Specifically, focusing on the ranges of 0.425–0.575 $\mu$m and 0.725–0.875 $\mu$m, the one using $\ell_1$-norm and TSV regularization (Figure 4i) is the closest to the input spectrum. These results indicate that the surface distribution and spectrum inferred by spin–orbit unmixing with $\ell_1$-norm and TSV regularization are superior to other regularizations.

Let us also note the solution inferred by spin–orbit unmixing with trace norm. The inferred surface distributions (Figures 4(g)–(i)) capture the features of continuous surfaces such as continents, but pixels with small values are noisier than those inferred with $\ell_1$-norm and TSV regularization. This may be due to the similarity of the vectors corresponding to the surface distributions of the end-members as a result of the low-rank matrix induced by using trace norm regularization. In this study, we employed only trace norm regularization in the inference. We can consider adding other regularizations or constraints, especially the nonnegative condition. The nonnegative condition leads to a map in which large parts are zero, as shown in Section 4.2 (spin–orbit unmixing with $\ell_1$-norm and TSV regularization), Kawahara (2020) (spin–orbit unmixing with Tikhonov regularization), and Kawahara & Fuji (2010) (spin–orbit tomography with Bounded Variable Least-Squares Solver including the nonnegative condition). Therefore, by including the nonnegative condition in spin–orbit unmixing with trace norm, we expect improvement in inferences of geography and spectrum. The application of trace norm regularization is subject to future investigation.

5. Application to Real Observed Data

In this section, we apply our method with $\ell_1$-norm and TSV regularization to real long-monitoring data of Earth as observed by DSCOVR/Earth Polychromatic Imaging Camera (EPIC; Jiang et al. 2018). Since 2015, DSCOVR has been continuously observing the dayside of Earth from the first Sun–Earth Lagrangian point (L1). Given that Earth’s rotation axis is tilted relative to its orbit, although it is not the same as that in the case
Figure 4. (a–c) Input surface distributions and (d, e, f, m) inferred solutions by spin–orbit unmixing with $\ell_1$-norm and TSV regularization using a cloudless Earth model, (g, h, i, n) trace norm regularization, and (j, k, l, o) Tikhonov regularization. The first, second, and third columns of panels (a)–(l) are vegetation, land, and ocean distributions, respectively. In the spectra in panels (m)–(o), the solid lines are the inferred spectra, and the dotted lines represent the input spectra. Panels (j), (k), (l), and (o) are based on Kawahara (2020).
of direct imaging, the observed data contain two-dimensional information about the planet surface. This allowed us to perform two-dimensional mapping (Fan et al. 2019). Kawahara (2020) inferred the surface distribution and unmixed spectra by applying spin–orbit unmixing with Tikhonov regularization to the DSCOVR data. In our experiment, we applied spin–orbit unmixing with the $\ell_1$-norm and TSV norm described in Section 3.1 to the DSCOVR data. Following the same setup as that in Kawahara (2020), we used one-quarter of the 2 yr data (i.e., one in each of the four bins) used in Fan et al. (2019). The observed wavelengths are the seven optical bands used in the EPIC instrument (0.388, 0.443, 0.552, 0.680, 0.688, 0.764, and 0.779 $\mu$m). There are strong oxygen B and A absorptions at 0.688 and 0.764 $\mu$m, respectively.

We selected the regularization parameters using the procedure described in Appendix E, as in Section 4.2. However, it is not possible to calculate MRS$A$ and CPR for actual exoplanet observations because the true surface distribution and reflection spectra are unknown. Therefore, only the mean residual and the normalized volume of the spectrum were used to determine the optimal parameters.

Figure 5 shows the mean residuals and normalized spectral volumes calculated by varying one of $\lambda_{\ell_1}, \lambda_{\text{TSV}},$ and $\lambda_X$. We selected $\lambda_{\ell_1} = 10^{-2}, \lambda_{\text{TSV}} = 10^{-2},$ and $\lambda_X = 10^{-7}$ as optimal values because the mean residual significantly increased at a range higher than these values, and the value of normalized spectral volume was sufficiently small ($\sim 10^{-20}$) at these values. As shown in Figure 6, as the calculation proceeds, the mean residual increases at a point. Therefore, we used the inferred solution at the epoch, where the mean residual is minimal.

Figure 6 shows the inferred spectra, color composite maps, and the map that excludes Component 0 using the aforementioned procedure (we assume that $N_k = 4$). In the unmixed spectra, the strong oxygen B and A absorption features were observed at 0.688 and 0.764 $\mu$m, respectively. As shown in Figure 7(b), we obtained the distribution of real clouds in the midlatitudes of the Southern Hemisphere, depicted by Component 0. In addition, there are few in the vicinity of the Sahara Desert. Therefore, we can interpret that Component 0 corresponds to the cloud. Kawahara (2020) using Tikhonov regularization resulted in patchy cloud distributions (Figure 8), but these results were inconsistent with the real distribution. In contrast, we were able to obtain more continuous maps of clouds than that from Tikhonov regularization. Note that real clouds are also distributed in the midlatitude zone of the Northern Hemisphere (described in Figure 6 in Kawahara & Masuda 2020), but we were unable to retrieve the distribution. This may be due to the degeneracy with the continents.

The map that excludes Component 0 (Figure 7(c)) accurately resembles the real continental distribution. The structure of South America and Australia, depicted by Components 2 and 3, respectively, is consistent with Aizawa et al. (2020; single-band mapping using sparse modeling). Component 1 depicts the geography of the ocean. The spectrum of Component 1 also reasonably reproduces that of the ocean. Hence, we can interpret that Component 1 corresponds to the ocean.

On the other hand, Components 2 and 3 seem to degenerate. However, we can see that the geographical features of North America and Australia are depicted by Component 2, and those of the Sahara Desert and Chile by Component 3. Furthermore, South America and Eurasia are depicted by both Components 2 and 3. This is probably because these continents contain both soil and vegetation on the surface. For unmixed spectra,
Component 2 exhibits larger values at 0.688 and 0.764 μm than Component 3, but smaller for 0.779 μm, that is, Component 3 appears redder than Component 2. Additionally, the increase at 0.688 and 0.764 μm might be interpreted that the spectrum of Component 2 captures the red edge of vegetation, although the strong oxygen absorption bands make the interpretation difficult. Thus, we can interpret that Component 2 corresponds to vegetation and Component 3 to soil or sands.

Let us also note that the South Pole should be depicted as ice, namely, Component 0, but it was depicted as Component 2 in Figure 7 and not visible in Figure 8. This inconsistency may be due to the low observational weights on the poles of DSCOVR (presented as Figure 3(b) in Aizawa et al. 2020). When compared with the inferred map obtained using Tikhonov regularization in Kawahara (2020) (Figure 8), the continents are better separated from each other in the map obtained using ℓ₁-norm and TSV regularization, especially for the Arabian Sea and North Atlantic Ocean.

6. Conclusion

In this study, we introduced sparse modeling (ℓ₁-norm and TSV regularization) to spin–orbit unmixing for the global mapping of planetary surfaces. For this purpose, we combined and improved the methods proposed by Aizawa et al. (2020) and Kawahara (2020) and modified the method proposed by Fan et al. (2019). Test calculations on a cloudless toy model of Earth yielded surface distributions with sparsity and continuity. The inferred unmixed spectra were closer to the input model than those inferred by Kawahara (2020). Applying our method to real observation data of Earth obtained by DSCOVR, we also found that the surface distributions and spectra were reasonably recovered by the current method. We concluded that sparse modeling provides better inferences of the surface distribution and unmixed spectra than the method based on Tikhonov regularization.

This study can be extended in several ways. In this study, we focused on the ℓ₁-norm and TSV regularization, which prefers sparsity and continuity. However, other choices of regularizations for surface distributions and spectrum can be considered. Furthermore, another type of sparse modeling based on matrices was proposed in previous studies (e.g., Candès et al. 2011), and different types of volume regularization in remote sensing can also be used (e.g., Ang & Gillis 2019). Additionally, we assumed the surface distribution of the end-member as static, but we should also consider the dynamical motion of surfaces, especially for clouds. Recently, Kawahara...
& Masuda (2020) developed dynamic spin–orbit tomography to retrieve the geometry and surface maps in a single band using Tikhonov regularization, and we can extend their method based on sparse modeling. Ultimately, we might be able to combine dynamical mapping (Kawahara & Masuda 2020) and spectral unmixing (Kawahara 2020) into dynamic spin–orbit unmixing to solve the dynamical motions of planetary surfaces. These issues will be addressed in future research.

The authors are indebted to the DSCOVR team for making the data publicly available. We are grateful to Siteng Fan and Yuk L. Yung for providing the processed light curves and their geometric kernel from the DSCOVR data set. We are also grateful to Kento Masuda and Shiro Ikeda for insightful discussions. We would also like to thank the anonymous reviewer for an attentive reading and fruitful suggestions. This study was supported by JSPS KAKENHI grant Nos. JP18H04577, JP18H01247, JP20H00170, JP21H04998 (H.K.), JP22000005, JP15H02063, and JP18H05442 (M.T.). A.K. was also supported by JST SPRING, grant No. JPMJSP2108. This study was supported by the JSPS Core-to-Core Program Planet and SATELLITE Research from the Astrobiology Center (H.K.).

Appendix A
Spectral Unmixing

In this appendix, we review spectral unmixing, which originates from remote sensing techniques. First, we consider hyperspectral images, which are targets of spectral unmixing. A hyperspectral image has dimensions in the spatial direction and also in the wavelength direction (Bioucas-Dias et al. 2013). Each pixel in the image contains multiple components (e.g., vegetation, land, and ocean). While we term it a pure pixel when a pixel contains only a single component, we term it a mixed pixel when it contains multiple components owing to the observational resolution, and each component is termed an end-member. Decomposition of the observed image into the spectra of the end-members and their abundance is termed as spectral unmixing.

We now consider spectral unmixing with NMF (Paatero & Tapper 1994). Let $N_j$, $N_c$, and $N_l$ denote the number of pixels of the image, end-members, and wavelengths of observation, respectively. The hyperspectral image obtained from the observation is $M \in \mathbb{R}^{N_j \times N_l}$, where $M_{j,l} = m(r_j, \lambda_l)$ represents the observational data at position $r_j$ and wavelength $\lambda_l$. The linear mixing model is expressed as follows:

$$M = AX,$$

where $A \in \mathbb{R}^{N_c \times N_j}$ denotes the surface distribution matrix and $X \in \mathbb{R}^{N_l \times N_j}$ denotes the end-member matrix. Here $a_{ij}$, which denotes the $k$th column vector of $A$, denotes the surface distribution of the $k$th end-member, namely, an abundance of $A_{ik} = a_{ik}(r_j)$; $x_k$, which denotes the $k$th column vector of $X^*$, denotes the reflection spectrum of the $k$th end-member $X_{kl} = x_k(\lambda_l)$. The problem that is described by matrices, such as spectral unmixing, is expressed as an optimization problem:

$$\min_{A,X} \frac{1}{2} ||M - AX||^2_F.$$

where $|| \cdot ||_F$ denotes the Frobenius norm defined as

$$||Y||_F := \sqrt{\sum \sum Y_{ij}^2}.$$

The optimization problem of adding the constraint that the entries of each matrix are nonnegative is NMF:

$$\min_{A,X} \frac{1}{2} ||M - AX||^2_F$$

subject to $A_{ik} \geq 0$, $X_{kl} \geq 0$.

On the other hand, we can generate matrices $A'$ and $X'$ using the regular matrix $P \in \mathbb{R}^{N_j \times N_j}$ in Equation (A1):

$$M = AX = A'X' ,$$

$$A' = AP ,$$

$$X' = P^{-1}X .$$

Thus, in general, $A$ and $X$ that satisfy Equation (A1) are not unique. Additionally, NMF is known to be NP-hard; thus, it is difficult to determine the optimal solution. To address the above problems, it is necessary to add appropriate constraints or regularization terms:

$$\min_{A,X} \frac{1}{2} ||M - AX||^2_F + R(A, X)$$

subject to $A_{ik} \geq 0$, $X_{kl} \geq 0$.

where $R(A, X)$ denotes the regularization term. With respect to NMF, using simplex volume minimization as a regularization term can reproduce the high-resolution spectrum components (Craig 1994; Lin et al. 2015; Ang & Gillis 2019; Fu et al. 2019). In the simplex volume regularization, $\det(XX^*)/(N_j!)$, the volume of an $(N_j - 1)$-simplex in $(N_j - 1)$-dimensional space with vertices $\{x_1, x_2, \ldots, x_{N_j}\}$, is used as the regularization term:

$$\min_{A,X} \frac{1}{2} ||M - AX||^2_F$$

$$+ \lambda X \frac{1}{2} \det(XX^*) \text{subject to } A_{ik} \geq 0, X_{kl} \geq 0 ,$$

where $\lambda X$ denotes the regularization parameter.

Appendix B
Optimization of a Nondifferentiable Function

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a function. The optimization problem of obtaining a solution $x \in \mathbb{R}^n$ that minimizes $f$ under constraint $x \in S \subset \mathbb{R}^n$ is expressed as follows:

$$\min_{x \in \mathbb{R}^n} f(x) \text{subject to } x \in S \text{(constrained problem)}.$$

Function $f$ is termed the objective function. Specifically, when $S = \mathbb{R}^n$, the optimization problem can be reexpressed as follows:

$$\min_{x \in \mathbb{R}^n} f(x) \text{(unconstrained problem)}.$$

In the following, we consider the objective function $f$ to be a convex function (Definition 1). Optimization problems for convex functions are extensively studied owing to their tractable properties (e.g., Rockafellar 1970).
If $f$ is a differentiable function, then the update formula for the gradient descent method, which is the simplest update method, can be provided as follows:

$$x_{i+1} = x_i - \gamma \nabla f(x_i),$$  \hspace{1cm} (B3)

where $\gamma > 0$ denotes the parameter indicating the step size.

### B.1. Proximal Point Algorithm

If $f$ is a nondifferentiable function, then we cannot use the gradient descent method because $\nabla f$ does not exist. An algorithm used to solve this problem is the proximal point algorithm (Kanamori et al. 2016). Let $\psi$ be a closed proper convex function (Definitions 3, 8) that is not necessarily differentiable. We consider the following optimization problem:

minimize $\psi(x)$.

(B4)

The update formula for the proximal point algorithm is expressed as

$$x_{i+1} = \text{prox}(x_i|\gamma \psi) := \arg\min_{y \in \mathbb{R}^n} \left( \psi(y) + \frac{1}{2\gamma} ||y - x_i||^2 \right),$$  \hspace{1cm} (B5)

where $\text{prox}(x|\gamma \psi)$ denotes the proximal operator of $\psi$ (Definition 18), and its value is unique for any $x$ (Proposition 19).

Now, we consider the following function for $\psi$:

$$\psi_\gamma(x) := \min_{y \in \mathbb{R}^n} \left( \psi(y) + \frac{1}{2\gamma} ||y - x||^2 \right).$$

(B6)

Subsequently, the function is expressed as

$$\psi_\gamma(x) = \frac{1}{2\gamma} ||x||^2 - \frac{1}{\gamma} \max_{y \in \mathbb{R}^n} \left( x^\top y - \psi(y) - \frac{1}{2} ||y||^2 \right),$$

(B7)

$$= \frac{1}{2\gamma} ||x||^2 - \frac{1}{\gamma} \left( \gamma \psi^*(y) + \frac{1}{2} ||y||^2 \right)^*,$$

(B8)

where $^*$ is the conjugate function (Definition 9). Given that $\gamma \psi^* + (1/2)||y||^2$ is a 1-strongly convex function (Theorem 6), its conjugate function is a 1-smooth function (Theorem 12). Hence, $(\gamma \psi^* + (1/2)||y||^2)^*$ is differentiable and $\psi_\gamma(x)$ is differentiable. We term $\psi_\gamma(x)$ the Moreau envelope of $\psi$, which smoothen $\psi$. The gradient of $\psi_\gamma(x)$ is

$$\nabla \psi_\gamma(x) = \frac{1}{\gamma} x - \frac{1}{\gamma} \arg\max_{y \in \mathbb{R}^n} \left( x^\top y - \gamma \psi(y) - \frac{1}{2} ||y||^2 \right)$$

(Corollary 15)

$$= \frac{1}{\gamma} x - \frac{1}{\gamma} \arg\min_{y \in \mathbb{R}^n} \left( \psi(y) + \frac{1}{2\gamma} ||y - x||^2 \right),$$

(B9)

and then we have

$$\text{prox}(x|\gamma \psi) = x - \gamma \nabla \psi_\gamma(x).$$

(B10)

Hence, the proximal point algorithm smoothens the objective function prior to applying the gradient descent method.

### B.2. Proximal Gradient Method

Let $f$ be a proper convex function that is differentiable, and let $\psi$ be a proper convex function that is not necessarily differentiable. We consider the following optimization problem:

$$\minimize_{x \in \mathbb{R}^n} f(x) + \psi(x).$$

(B13)

An algorithm used to solve this problem is the proximal gradient method (Kanamori et al. 2016). Its first updates

$$w_i = x_i - \gamma \nabla f(x_i)$$

(B14)

using the gradient descent method for $f$ and then updates

$$x_{i+1} = \text{prox}(w_i|\gamma \psi)$$

(B15)

using the proximal point algorithm for $\psi$. These are summarized as

$$x_{i+1} = \text{prox}(x_i - \gamma \nabla f(x_i)|\gamma \psi).$$

(B16)

Next, we consider the optimization problem for a differentiable proper convex function $f$ with nonnegative constraints as follows:

$$\minimize_{x \in \mathbb{R}^n} f(x) \text{ subject to } x_j \geq 0 \quad (j = 1,\ldots,n).$$

(B17)

The problem is equivalent to

$$\minimize_{x \in \mathbb{R}^n} f(x) + \delta_+(x),$$

(B18)

where $\delta_+$ denotes the indicator function of a nonnegative set defined as

$$\delta_+(x) := \begin{cases} 0 & (x \geq 0) \\ \infty & \text{otherwise} \end{cases}$$

(B19)

where $x \geq 0$ denotes $x \in \{x \in \mathbb{R}^n | x_j \geq 0 \quad (j = 1,\ldots,n)\}$. Given that $\delta_+$ is a closed proper convex function (Proposition 17), the proximal operator of $\delta_+$ can be defined:

$$\text{prox}(x|\delta_+) := \arg\min_{y \geq 0} \left( \delta_+(y) + \frac{1}{2\gamma} ||y - x||^2 \right)$$

(B20)

$$= \arg\min_{y \geq 0} ||y - x||^2$$

(B21)

$$= \max(x, 0),$$

(B22)

where max denotes the element-wise maximum. The update formula of the proximal gradient method for problem (B17) is written as

$$x_{i+1} = \max\{x_i - \gamma \nabla f(x_i), \mathbf{0}\}.$$  \hspace{1cm} (B23)

### Appendix C

**Sparse Optimization Problem**

Sparse modeling is a technique that extracts and analyzes low-dimensional information to explain high-dimensional data. We consider a method to infer the sparse solution to optimization problems for a differentiable proper convex
function $f$:

$$\minimize_{x \in \mathbb{R}^n} f(x). \quad (C1)$$

### C.1. $\ell_1$-norm Regularization

The sparsity of a solution implies that most of its elements are zero. We then infer with the constraint to reduce the number of nonzero elements. A straightforward method to infer a sparse solution involves solving the following optimization problem:

$$\minimize_{x \in \mathbb{R}^n} f(x) \text{subject to } ||x||_0 \leq C, \quad (C2)$$

where $||x||_0 := \# \{ j \mid x_j \neq 0 \}$ denotes the $\ell_0$-norm of $x$ and $C$ denotes a parameter. However, solving Equation (C2) incurs a huge computational cost because function values must be calculated continuously while changing the value of $||x||_0$. A method of optimization using the $\ell_1$-norm instead of the $\ell_0$-norm is proposed as follows (Tibshirani 1996):

$$\minimize_{x \in \mathbb{R}^n} f(x) \text{subject to } ||x||_1 \leq C, \quad (C3)$$

where $||x||_1 := \sum_{i=1}^{n} |x_i|$ denotes the $\ell_1$-norm. A sparse solution can be inferred by using this method. The $\ell_1$-norm is an approximation of the $\ell_0$-norm because $||x||_1$ is a convex hull of $||x||_0$ in $[-1, 1]^n$. Moreover, Equation (C3) is equivalent to

$$\minimize_{x \in \mathbb{R}^n} f(x) + \lambda ||x||_1, \quad (C4)$$

where $\lambda$ denotes a parameter (Tomikawa 2015). We can solve problem (C4) using the proximal gradient method because $||x||_1$ is a nondifferentiable proper convex function.

### C.2. Total Variation and Total Squared Variation

Another example of sparse modeling is the total variation (TV) regularization, defined as

$$\minimize_{x \in \mathbb{R}^n} f(x) + \lambda ||x||_{TV}, \quad (C5)$$

where $\lambda$ denotes the regularization parameter. We then define matrix $N$ to represent adjacent pixels:

$$N_{ij} := \begin{cases} 1 & \text{if $ith$ and $jth$ pixels are adjacent} \\ 0 & \text{(otherwise)}. \end{cases} \quad (C6)$$

The TV regularization term is written using $N$ as

$$||x||_{TV} := \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2} N_{ij} |x_i - x_j|. \quad (C7)$$

TV regularization minimizes the difference between the values of neighboring pixels. This is expected to smooth the values of the neighboring pixels and reduce noise in the solution.

Moreover, the TSV (Kuramochi et al. 2018) is expressed as an extension of TV regularization:

$$\minimize_{x \in \mathbb{R}^n} f(x) + \lambda ||x||_{TSV}, \quad (C8)$$

$$||x||_{TSV} := \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2} N_{ij} (x_i - x_j)^2. \quad (C9)$$

TSV regularization allows us to infer solutions with smooth boundaries in addition to the effects of TV regularization.

consider the discretized distribution using HEALPix, the TSV regularization term can be reexpressed as

$$||x||_{TSV} = x^T \tilde{N} x, \quad (C10)$$

$$\tilde{N} := 8I - N, \quad (C11)$$

because

$$||x||_{TSV} = \sum_{i} \sum_{j} \frac{1}{2} N_{ij} x_i^2 + \sum_{j} \sum_{i} \frac{1}{2} N_{ij} x_j^2 - \sum_{i} \sum_{j} N_{ij} x_i x_j \quad (C12)$$

$$= \sum_{i} \sum_{j} \frac{1}{2} N_{ij} x_i^2 + \sum_{j} \sum_{i} \frac{1}{2} N_{ij} x_j^2 - \sum_{i} \sum_{j} N_{ij} x_i x_j \quad (C13)$$

$$= \sum_{i} \sum_{j} N_{ij} x_i^2 - \sum_{i} \sum_{j} N_{ij} x_i x_j (N_{ij} = N_{ji}) \quad (C14)$$

$$= \sum_{i} 8x_i^2 - \sum_{i} \sum_{j} N_{ij} x_i x_j \quad (C15)$$

(eight neighboring pixels per pixel)

$$= \sum_{i} 8x_i \sum_{j} \delta_{i j} x_j - \sum_{i} \sum_{j} N_{ij} x_i x_j \quad (C16)$$

$$= \sum_{i} \sum_{j} x_i (8 \delta_{i j} - N_{ij}) x_j \quad (C17)$$

$$= x^T (8I - N) x. \quad (C19)$$

### C.3. Trace Norm Regularization

A type of sparse modeling that uses the structure of matrices is trace norm regularization. First, the singular value decomposition of matrix $X$ is written as $X = U \Sigma V^T$, where $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{r \times r}$ are orthogonal matrices, $\Sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix, and $r = \min \{m, n\}$. When $\Sigma = \text{diag}(\sigma_1(X), \ldots, \sigma_r(X))(\sigma_1(X) > \sigma_2(X) > \cdots > \sigma_r(X))$, $\sigma_j(X)$ ($j = 1, \ldots, r$) is termed as the $j$th singular value. We define the trace norm of the matrix $X$ as follows:

$$||X||_\text{Tr} := \sum_{j=1}^{r} \sigma_j(X). \quad (C20)$$

Let $\sigma(X)$ be a singular value vector of $X$, defined as $(\sigma(X))_j := \sigma_j(X)$; then, the trace norm of $X$ is reexpressed as

$$||X||_\text{Tr} = ||\sigma(X)||_1. \quad (C21)$$

Hence, we can infer the matrix with a sparse singular value vector by solving the optimization problem with the trace norm as the regularization term:

$$\minimize_{X \in \mathbb{R}^{m \times n}} f(X) + \lambda ||X||_\text{Tr}, \quad (C22)$$

where $\lambda$ denotes the regularization term. The number of nonzero singular values of a matrix $X$ is equal to the rank of $X$, and thus trace norm regularization allows us to infer a low-rank matrix.

### Appendix D

**Spin–Orbit Unmixing with Tikhonov Regularization**

For comparison with sparse modeling, we also consider spin–orbit unmixing with Tikhonov regularization for
geography $A$ and volume regularization for $X$, minimize $Q_{A,k}$ subject to $A_{jk} \geq 0, X_{kl} \geq 0$, \hspace{1cm} (D1)

\[ Q_{A,k} := \frac{1}{2} \| D - W A X \|_F^2 + \frac{\lambda_A}{2} \| A \|_F^2 + \frac{\lambda_X}{2} \det(X X^T). \] \hspace{1cm} (D2)

By rewriting $Q_{A,k}$ as the quadratic form of $a_k$, the $k$th column vector of $A$, we obtain

\[ Q_{A,k} = \frac{1}{2} a_k^T (L_A + T_A) a_k - I_A a_k + (\text{constant for } a_k), \] \hspace{1cm} (D3)

where $L_A := x_k^T x_k W^T W$, $T_A := \lambda_A I$, $I_k := W^T \Delta x_k$, and $\Delta$ denotes a matrix defined as $\Delta_{il} := D_{il} - \sum_{j=1}^N W_{ij} A_{jk} x_{jl}$.

Furthermore, from $Q_{A,k}$ to the quadratic form of $x_{kl}$, the $k$th column vector of $X^T$, we obtain

\[ Q_{X,k} = \frac{1}{2} x_{kl}^T (L_X + D_X) x_{kl} - I_X x_{kl} + (\text{constant for } x_{kl}), \] \hspace{1cm} (D4)

where $L_X := \| W a_k \|_2^2 I$, $D_X := \lambda_X \det(\tilde{X}_k \tilde{X}_k^T) (I - \tilde{X}_k (\tilde{X}_k \tilde{X}_k^T)^{-1} \tilde{X}_k)$, $I_k := \Delta_D W a_k$, $I$ denotes an identity matrix, and $\tilde{X}_k$ denotes a submatrix of $X$ with the $k$th row removed.

The subproblem to solve optimization problem (D1) is expressed as

minimize $q_{A,Tik}$ subject to $(a_k)_j \geq 0$ $(j = 1, \ldots, N_j)$, \hspace{1cm} (D5)

minimize $q_{X,Tik}$ subject to $(x_{kl})_l \geq 0$ $(l = 1, \ldots, N_j)$, \hspace{1cm} (D6)

\[ q_{A,Tik} := \frac{1}{2} a_k^T (L_A + T_A) a_k - I_A a_k, \] \hspace{1cm} (D7)

\[ q_{X,Tik} := \frac{1}{2} x_{kl}^T (L_X + D_X) x_{kl} - I_X x_{kl}. \] \hspace{1cm} (D8)

Given that Equations (D5) and (D6) are optimization problems with nonnegative constraints for differentiable objective functions (B17), they can be solved using the proximal gradient method. In general, for optimization problem

\[ \text{minimize } q(w) \text{ subject to } w_j \geq 0 \ (j = 1, \ldots, N_j), \hspace{1cm} (D9) \]

\[ q(w) := \frac{1}{2} w^T L w - I w, \] \hspace{1cm} (D10)

we have $\nabla q(w) = L w - I$; thus, the update formula of the proximal gradient method for (D9) is written as

\[ w_{t+1} = \max \{ w_t - \gamma (L w_t - I), 0\}. \] \hspace{1cm} (D11)

Kawahara (2020) performed optimization by solving subproblems (D5) and (D6) at each $k$. FISTA (Beck & Teboulle 2009b), which is a proximal gradient method using Nesterov’s acceleration method (Nesterov 2003), is used to solve each subproblem. Nesterov’s acceleration method increases the convergence speed, although the function value does not necessarily decrease monotonically. A method to solve this involves using the restart method (O’donoghue & Candes 2015), which restarts Nesterov’s acceleration method when the function value does not decrease. Hence, the optimization algorithm for solving Equation (D1) is expressed as follows:

**Algorithm 3.** Spin–orbit unmixing with Tikhonov regularization

| Initialization: nonnegative matrices $A_0$ and $X_0$ for $n$ in $(1, N_0)$ do for $k$ in $(1, N_j)$ do Update $x_k$ using FISTA with the restart method Update $a_k$ using FISTA with the restart method end for end for

**Appendix E**

**Evaluation of Inferred Solutions**

We present the evaluation measures used as criteria to select parameters $\lambda_0$, $\lambda_{TSV}$, and $\lambda_X$ in optimization problem (21). The first is the mean residual

\[ \text{mean residual} := \frac{1}{D} \sqrt{\frac{\| D - WAX \|_F^2}{N_N N_j}}, \] \hspace{1cm} (E1)

which denotes the difference between the inferred model and observed data. The first row in Figure 9 shows mean residuals for each parameter.

The second row of Figure 9 shows $\det(X'X'^T)$, which corresponds to the volume of a simplex with each column vector of the normalized spectrum $X'$, where $X'$ is defined as

\[ X'_il := \frac{\hat{X}_il}{\sum_{j=1}^N \hat{X}_jl} \hspace{1cm} (k = 1, \ldots, N_k, \ l = 1, \ldots, N_l). \] \hspace{1cm} (E2)

Given that the mean residuals tend to increase as $\det(X'X'^T)$ decreases from Figure 9, a trade-off exists between these two values in spin–orbit unmixing with the $\ell_1$-norm and TSV regularization. This relationship is also observed in spin–orbit unmixing with Tikhonov regularization (Kawahara 2020).

The MRSA is defined as an evaluation measure that directly compares the inferred spectra with the true values:

\[ \text{MRSA}(x, y) := \frac{1}{\pi} \cos^{-1}\left\{ \frac{(x - \tau 1)(y - \tau 1)}{\|x - \tau 1\|_2 \|y - \tau 1\|_2} \right\}. \] \hspace{1cm} (E3)

Because $\pi \text{MRSA}(x, y)$ is the angle between $(x - \tau 1)/\|x - \tau 1\|_2$ and $(y - \tau 1)/\|y - \tau 1\|_2$, we obtain that $0 \leq \pi \text{MRSA}(x, y) \leq \pi$. Thus, we get

\[ 0 \leq \text{MRSA}(x, y) \leq 1. \] \hspace{1cm} (E4)

Specifically, if $x = y$, $\text{MRSA}(x, y) = 0$. We used the average of the inferred spectra at each end-member and then compared with the true spectra as the evaluation measure of the model:

\[ \text{MRSA} := \frac{1}{N_k} \sum_{k=1}^N \text{MRSA}(x_k, x_{k,\text{true}}). \] \hspace{1cm} (E5)

The third row of Figure 9 shows $\text{MRSA}$ for each parameter.

Furthermore, we consider the CPR as an evaluation measure that directly compares the surface distribution with the true value. The classification of the $j$th pixel of the inferred surface
distribution is

\[ c_j(A) = \arg \max_{k \in \{1, \ldots, N_k\}} (a_k)_j, \quad (E6) \]

namely, \( c_j(A) = k \) denotes the assignment of the \( j \)th pixel of the surface distribution to the \( k \)th end component. We then define CPR as

\[ \text{CPR} = \frac{1}{N_j} \# \{ j | c_j(\hat{A}) = c_j(A_{\text{true}}) \}. \quad (E7) \]

The fourth row of Figure 9 shows CPR for each parameter.

We select a regularization parameter that resulted in smaller mean residuals, \( \det(X'X^T) \), and \( \text{MRSA} \) and a larger CPR. Given the large rate of change in the value of \( \text{MRSA} \), as shown in Figure 9, we measured the value of each evaluation measure by changing the parameter around the local minimum value of \( \text{MRSA} \). Furthermore, \( \lambda_{T} = 10^{-3.5}, \lambda_{TSV} = 10^{-4}, \) and \( \lambda_{X} = 10^{2} \) when \( \text{MRSA} \) exhibits a local minimum. At these points, the mean residual and \( \det(X'X^T) \) values are sufficiently small, and the CPR is sufficiently large. Thus, we adopt \( \lambda_{T} = 10^{-3.5}, \lambda_{TSV} = 10^{-4}, \) and \( \lambda_{X} = 10^{2} \) under the assumption that the rating scale changes independently for each regularization parameter. Figure 10 is the same as Figure 9, but with trace norm. We adopt \( \lambda_{T} = 10^{2.1} \) and \( \lambda_{X} = 10^{3.3} \).
Figure 10. Same as Figure 9, but for trace norm. $\lambda_A = 10^{2.1}$ and $\lambda_X = 10^{3.3}$ are fixed in the first and second columns, respectively, and their values are calculated by changing the other parameter.
Appendix F
Theory of Convex Optimization

In this section, we review the theory of convex optimization, which is the basis of spin–orbit unmixing.

F.1. Basis of a Convex Function

Definition 1. Convex function, Rockafellar (1970, p. 23). Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{\pm \infty\}$ be a function. $f$ is said to be convex if

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$$  \hspace{1cm} (F1)

for any $x, y \in \mathbb{R}^d$ and any $\alpha \in [0, 1]$.

Definition 2. Effective domain, Rockafellar (1970, p. 23). Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{\pm \infty\}$ be a convex function. The effective domain of $f$ is as follows:

$$\text{dom } f := \{x \in \mathbb{R}^n | f(x) < +\infty\}. \hspace{1cm} (F2)$$

Definition 3. Proper convex function, Rockafellar (1970, p. 24). Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{\pm \infty\}$ be a convex function. $f$ is said to be proper if

$$\left\{ \begin{array}{l}
\text{for any } x \in \mathbb{R}^n, f(x) > -\infty \\
\text{there exists } x \in \mathbb{R}^n \text{ such that } f(x) < +\infty,
\end{array} \right. \hspace{1cm} (F3)$$

that is,

$$\left\{ \begin{array}{l}
f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \\
\text{dom } f \neq \emptyset.
\end{array} \right. \hspace{1cm} (F4)$$

Definition 4. $\mu$-strongly convex function, Rockafellar & Wets (2009, p. 565; Definition 12.58). Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. $f$ is said to be $\mu$-strongly convex if there exists $\mu > 0$ such that for any $x, y \in \text{dom } f$ and any $\alpha \in [0, 1]$,

$$\frac{\mu}{2} \alpha (1 - \alpha)\|x - y\|^2_i + f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y). \hspace{1cm} (F5)$$

Hence, a ($\mu$-)strongly convex function is strictly convex.

Theorem 5. Fukushima (2001, p. 58; Theorem 2.42). Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a ($\mu$-)strongly convex function. There uniquely exists a minimum value of $f$.

Theorem 6. Rockafellar & Wets (2009, p. 565). Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. The following statements are equivalent:

(i) $f$ is a $\mu$-strongly convex function.

(ii) $\tilde{f} := f(x) - \frac{\mu}{2}\|x\|^2_i$ is a proper convex function.

Proof. For any $x, y \in \text{dom } f$ and any $\alpha \in [0, 1]$, we have

$$\tilde{f}(\alpha x + (1 - \alpha)y) = f(\alpha x + (1 - \alpha)y) - \frac{\mu}{2} \alpha (1 - \alpha)\|x - y\|^2_i \hspace{1cm} (F6)$$

$$= \alpha \tilde{f}(x) + (1 - \alpha)\tilde{f}(y)$$

$$- \left\{ \alpha f(x) + (1 - \alpha)f(y) - \frac{\mu}{2} \alpha (1 - \alpha)\|x - y\|^2_i \right\}$$

$$- f(\alpha x + (1 - \alpha)y). \hspace{1cm} (F7)$$

(i) $\Rightarrow$ (ii) Suppose $f$ is $\mu$-strongly convex. Since

$$\alpha f(x) + (1 - \alpha)f(y) - \frac{\mu}{2} \alpha (1 - \alpha)\|x - y\|^2_i$$

$$- f(\alpha x + (1 - \alpha)y) \geq 0 \hspace{1cm} (F8)$$

by Definition 4, we obtain

$$\tilde{f}(\alpha x + (1 - \alpha)y) \leq \alpha \tilde{f}(x) + (1 - \alpha)\tilde{f}(y). \hspace{1cm} (F9)$$

Therefore, $\tilde{f}$ is convex. Furthermore, by the definition of $\tilde{f}$, we obtain the following:

$$\left\{ \begin{array}{l}
\tilde{f} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \\
\text{dom } \tilde{f} \neq \emptyset.
\end{array} \right. \hspace{1cm} (F10)$$

Hence, $\tilde{f}$ is proper convex.

(i) $\Rightarrow$ (ii) Suppose $\tilde{f}$ is proper convex. Since

$$\tilde{f}(\alpha x + (1 - \alpha)y) \leq \alpha \tilde{f}(x) + (1 - \alpha)\tilde{f}(y). \hspace{1cm} (F11)$$

by Definition 4, we obtain the following:

$$\alpha f(x) + (1 - \alpha)f(y) - \frac{\mu}{2} \alpha (1 - \alpha)\|x - y\|^2_i$$

$$- f(\alpha x + (1 - \alpha)y) \geq 0 \hspace{1cm} (F12)$$

which is equivalent to Definition 4. Hence, $f$ is $\mu$-strongly convex. \hfill \Box

Proposition 7. Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The following statements are true.

(i) If $g$ is a convex function, then $f + g$ is a convex function.

(ii) If $g$ is a $\mu$-strongly convex function and $\text{dom } f \cap \text{dom } g \neq \emptyset$, then $f + g$ is $\mu$-strongly convex.

Proof.

(i) Suppose $g$ is convex. For any $x, y \in \text{dom}(f + g) = \text{dom } f \cap \text{dom } g$ and any $\alpha \in [0, 1]$, we have

$$(f + g)(\alpha x + (1 - \alpha)y) = f(\alpha x + (1 - \alpha)y) + g(\alpha x + (1 - \alpha)y) \hspace{1cm} (F13)$$

$$\leq f(x) + (1 - \alpha)f(y) + g(x) + (1 - \alpha)g(y) \hspace{1cm} (F14)$$

(convexity of $f$ and $g$)

$$= \alpha (f + g)(x) + (1 - \alpha)(f + g)(y). \hspace{1cm} (F15)$$

Hence, $f + g$ is convex.

(ii) Suppose $g$ is $\mu$-strongly convex. The function $\tilde{g}(x) := g(x) - \frac{\mu}{2}\|x\|^2_i$ is proper convex according to Theorem 6. Given that $\text{dom } f \cap \text{dom } g \neq \emptyset$, $f + \tilde{g}$ is proper convex based on (i) of Proposition 7 and Definition 3. Moreover, we obtain the following:

$$f + \tilde{g}(x) = f(x) + g(x) - \frac{\mu}{2}\|x\|^2_i \hspace{1cm} (F17)$$

$$=(f + g)(x) - \frac{\mu}{2}\|x\|^2_i. \hspace{1cm} (F18)$$

Hence, $f + g$ is $\mu$-strongly convex by Theorem 6. \hfill \Box
Definition 8. Closed convex function, Rockafellar (1970, p. 52), Fukushima (1970, p. 51). Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. Function $f$ is said to be closed if $\{x \in \mathbb{R}^n | f(x) \leq c\}$ is a closed set for any $c \in \mathbb{R}$.

Definition 9. Conjugate function, Rockafellar (1970, p. 104). Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. The conjugate function $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ of $f$ is defined as follows:

$$f^*(x) := \sup_{y \in \mathbb{R}^n} (x^T y - f(y)). \quad (F19)$$

Theorem 10. Rockafellar (1970, p. 104). Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a closed proper convex function. For function $f^{**}$, which is a conjugate function of $f^*$, $f^{**} = f$ is true.

Definition 11. Smooth function, Kanamori et al. (2016, p. 14). Let $f: \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. $f$ is said to be smooth if there exists $\gamma > 0$ such that for any $x, y \in \mathbb{R}^n$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \gamma \|x - y\|_2. \quad (F20)$$

Theorem 12. Rockafellar & Wets (2009, p. 566 Proposition 12.60). Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a closed proper convex function. For any $\mu > 0$, the following statements are equivalent:

(i) $f$ is $\mu$-strongly convex function.
(ii) $f$ is $(1+\mu)$-smooth function.

Proposition 13. Rockafellar (1970, p. 242 Theorem 25.1). Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a function such that $\text{dom } f$ is an open convex set. When $f$ is differentiable in $\text{dom } f$, the following statements are equivalent:

(i) $f$ is convex function.
(ii) For any $x, y \in \text{dom } f$, $f(y) - f(x) \geq \nabla f(x)^T (y - x)$.

Proposition 14. Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a differentiable closed proper convex function. The following statements are equivalent:

(i) $g = \nabla f(x)$.
(ii) $x = \nabla f^*(g)$.
(iii) $f(x) + f^*(g) = g^T x$.

Proof. Let $\hat{y} = \text{argmax}_{y \in \mathbb{R}^n} (x^T y - f(y))$. Using (i) $\iff$ (iii) for $f^*$.

$$\hat{y} = \text{argmax}_{y \in \mathbb{R}^n} (x^T y - f(y)). \quad (F25)$$

Therefore, (ii) $\iff$ (iii) is true.

Corollary 15. Let $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a differentiable closed proper convex function. Then, we obtain the following:

$$\nabla f^*(x) = \text{argmax}_{y \in \mathbb{R}^n} (x^T y - f(y)). \quad (F26)$$

Proof. Suppose $\hat{y} = \text{argmax}_{y \in \mathbb{R}^n} (x^T y - f(y))$. Using $f^*(x) = x^T \hat{y} - f(\hat{y})$ by the definition of a conjugate function and Proposition 14, we obtain $\hat{y} = \nabla f^*(x)$.

F.2. Application for Optimization Problems

Definition 16. Indicator function, Rockafellar (1970, p. 28). Let $S \subset \mathbb{R}^n$ be a set. The indicator function $\delta_S$ is defined as follows:

$$\delta_S(x) := \begin{cases} 0 & (x \in S) \\ \infty & (x \notin S). \end{cases} \quad (F27)$$

Proposition 17. Fukushima (1970, p. 60, Theorem 2.43). If $S$ is a nonempty convex set, then $\delta_S$ is a proper convex function. If $S$ is also closed, then $\delta_S$ is a closed proper convex function.

Proof. Given that $S$ is a convex set, we obtain the following:

$$\alpha u + (1 - \alpha)v \in S \quad (F28)$$

for any $u, v \in S$ and $\alpha \in [0, 1]$. We consider an arbitrary $x, y \in \mathbb{R}^n$. If $x, y \in S$, we obtain $\alpha x + (1 - \alpha)y \in S$ for any $\alpha \in [0, 1]$. Thus, we obtain

$$0 = \delta_S(\alpha x + (1 - \alpha)y) \leq \alpha \delta_S(x) + (1 - \alpha)\delta_S(y) = 0. \quad (F29)$$

When at least one of $x$ or $y$ is not included in $S$, the function value becomes $\infty$. This also satisfies the definition of a convex function (Definition 1). Therefore, $\delta_S$ is a convex function. Furthermore, $\delta_S$ is a proper convex function based on the nonemptiness of $S$ and definition of $\delta_S$. Using Definition 8, the following statement is true:

$$\delta_S \text{ is a closed function } \iff \{x \in \mathbb{R}^n | \delta_S(x) \leq c\} \text{ is a closed set for any } c \in \mathbb{R} \quad (F30)$$

$$\iff \{x \in \mathbb{R}^n | x \in S\} \text{ is a closed set } \quad (F31)$$

$$\iff S \text{ is a closed set. } \quad (F32)$$

Therefore, if $S$ is a nonempty closed convex set, $\delta_S$ is a closed proper convex function.

Definition 18. Proximal Operator, Moreau (1965, p. 284). Let $\psi: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a closed proper convex function. The proximal operator of $\psi$ is defined as

$$\text{prox}(x|\psi) := \text{argmin}_{w \in \mathbb{R}^n} \left( \psi(w) + \frac{1}{2} \|w - x\|_2^2 \right). \quad (F33)$$

Proposition 19. Moreau (1965, p. 284). Let $\psi: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. Then, the value of $\text{prox}(x|\psi)$ is unique.
Proof. Given that function $\|\cdot\|^2_2$ is a strongly convex function, function $\psi(w) + (1/2)\|w - x\|^2_2$ is also a strongly convex function by Proposition 7. Hence, the value of $\text{prox}(x|\psi)$ is unique based on Theorem 5.

ORCID iDs
Atsuki Kuwata https://orcid.org/0000-0002-3244-7136
Hajime Kawahara https://orcid.org/0000-0003-3309-9134
Masataka Aizawa https://orcid.org/0000-0001-8877-4497
Takayuki Kotani https://orcid.org/0000-0001-6181-3142
Motohide Tamura https://orcid.org/0000-0002-6510-0681

References
Aizawa, M., Kawahara, H., & Fan, S. 2020, ApJ, 896, 22
Ang, A. M. S., & Gillis, N. 2019, IJSTA, 12, 4843
Asensio Ramos, A., & Pallé, E. 2021, A&A, 646, A4
Baldridge, A. M., Hook, S., Grove, C., & Rivera, G. 2009, RSEnv, 113, 711
Beck, A., & Teboulle, M. 2009a, ITIP, 18, 2419
Beck, A., & Teboulle, M. 2009b, SIIMS, 2, 183
Bioucas-Dias, J. M., Plaza, A., Camps-Valls, G., et al. 2013, IEEE Geosci. Remote Sens. Mag., 1, 6
Candes, E. J., Li, X., Ma, Y., & Wright, J. 2011, J. Assoc. Comput. Mach., 58, 1
Cowan, N. B., & Strait, T. E. 2013, ApJL, 765, L17
Craig, M. D. 1994, ITGRS, 32, 542
Fan, S., Li, C., Li, J.-Z., et al. 2019, ApJL, 882, L1
Farr, B., Farr, W. M., Cowan, N. B., Haggard, H. M., & Robinson, T. 2018, AJ, 156, 146
Ford, E. B., Seager, S., & Turner, E. 2001, Nat, 412, 885
Fu, X., Huang, K., Sidirooulos, N. D., & Ma, W.-K. 2019, ISPM, 36, 59
Fu, X., Ma, W.-K., Huang, K., & Sidiroopoulos, N. D. 2015, ITSP, 63, 2306
Fujii, Y., & Kawahara, H. 2012, ApJ, 755, 101
Fujii, Y., Lustig-Yaeger, J., & Cowan, N. B. 2017, AJ, 154, 189
Fukushima, M. 2001, Fundamentals of Nonlinear Optimization (Tokyo: Asakura Shoten)
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Jiang, J. H., Zhai, A. J., Herman, J., et al. 2018, AJ, 156, 26
Kanamori, T., Suzuki, T., Takeuchi, I., & Sato, I. 2016, Continuous Optimization for Machine Learning (Tokyo: Kodansha Scientific)
Kawahara, H. 2020, ApJ, 894, 58
Kawahara, H., & Fujii, Y. 2010, ApJ, 720, 1333
Kawahara, H., & Fujii, Y. 2011, ApJL, 739, L62
Kawahara, H., & Masuda, K. 2020, ApJ, 900, 48
Kim, J., He, Y., & Park, H. 2014, J. Global Optim., 58, 285
Kuramochi, K., Akiyama, K., Ikeda, S., et al. 2018, ApJ, 858, 56
Lin, C.-H., Ma, W.-K., Li, W.-C., Chi, C.-Y., & Ambikapathi, A. 2015, ITGRS, 53, 5530
Luger, R., Agol, E., Bartolić, F., & Foreman-Mackey, D. 2021, arXiv:2103.06275
Luger, R., Bedell, M., Vanderspek, R., & Burke, C. J. 2019, arXiv:1903.12182
Lustig-Yaeger, J., Meadows, V. S., Tovar Mendoza, G., et al. 2018, AJ, 156, 301
McLinden, C., McConnell, J., Griffioen, E., McElroy, C., & Pfister, L. 1997, JGRD, 102, 18801
Moreau, J.-J. 1965, Bull. Soc. Math. Fr., 93, 273
Nesterov, Y. 2003, Introductory Lectures on Convex Optimization: A Basic Course, 87 (Boston, MA: Springer)
O’donoghue, B., & Candès, E. 2015, Found. Comput. Math., 15, 715
Paatero, P., & Tapper, U. 1994, Environmetrics, 5, 111
Rockafellar, R. T. 1970, Convex Analysis, 11 (Princeton, NJ: Princeton Univ. Press)
Rockafellar, R. T., & Wets, R. J.-B. 2009, Variational Analysis, 317 (Berlin: Springer)
Schwartz, J. C., Sekowski, C., Haggard, H. M., Pallé, E., & Cowan, N. B. 2016, MNARS, 457, 926
Tibshirani, R. 1996, JRSSB, 58, 267
Tomoiaga, R. 2015, Machine Learning with Sparsity Inducing Regularizations (Tokyo: Kodansha Scientific)