MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Holonomic Gradient Descent and its Application to Fisher-Bingham Integral

Tomonari SEI, Nobuki TAKAYAMA, Akimichi TAKEMURA, Hiromasa NAKAYAMA, Kenta NISHIYAMA, Masayuki NORO and Katsuyoshi OHARA

METR 2010–15 May 2010

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

Holonomic Gradient Descent and its Application to Fisher-Bingham Integral

Tomonari Sei, Nobuki Takayama, Akimichi Takemura ;
Hiromasa Nakayama, Kenta Nishiyama, Masayuki Noro, Katsuyoshi Ohara

May 28, 2010

We give a new algorithm to find local maximum and minimum of a holonomic function and apply it for the Fisher-Bingham integral on the sphere S^n, which is used in the directional statistics. The method utilizes the theory and algorithms of holonomic systems.

1 Introduction

The gradient descent is a general method to find a local minimum of a smooth function $f(z_1, \ldots, z_d)$. The method utilizes the observation that $f(p)$ decreases if one goes from a point $z = p$ to a "nice" direction, which is usually $-(\nabla f)(p)$. As textbooks on optimizations present (see, e.g., [14]), we have a lot of achievements on this method and its variations.

We suggest a new variation of the gradient descent, which works for real valued *holonomic functions* $f(z_1, \ldots, z_d)$ and is a d-variable generalization of Euler's method for solving ordinary differential equations numerically and finding a local minimum of the function. We show an application of our method to directional statistics. In fact, it is our motivating problem to develop the new method.

A function f is called a holonomic function, roughly speaking, if f satisfies a system of linear differential equations

$$\ell_1 \bullet f = \ldots = \ell_r \bullet f = 0, \quad \ell_i \in D \tag{1}$$

whose solutions form a finite dimensional vector space. Here, D is the ring of differential operators with polynomial coefficients $\mathbf{C}\langle z_1, \ldots, z_d, \partial_1, \ldots, \partial_d \rangle$, $\partial_i = \partial/\partial z_i$.

Let us give a rigorous definition of holonomic function. A multi-valued analytic function f defined on $\mathbf{C}^d \setminus V$ with an algebraic set V is called a *holonomic function* if there exists a set of linear differential operators $\ell_i \in D$ annihilating f as (1) such that the left ideal generated by $\{\ell_1, \ldots, \ell_r\}$ in D is a *holonomic ideal* (see [13]). The function f is called real valued when a branch of f takes real values on a connected component of $(\mathbf{C}^d \setminus V) \cap \mathbf{R}^d$.

We give an equivalent definition of holonomic function without the notion of the holonomic ideal ([16], [11], [13]). A multi-valued analytic function f is called a holonomic function if f satisfies linear ordinary differential equations with polynomial coefficients for all variables $z_1, \ldots, z_d$. In other words, the function f satisfies a set of ordinary differential equations

$$\sum_{k=0}^{r_i} a_k^i(z_1, \ldots, z_d)\partial_i^k \bullet f = 0, \quad a_k^i \in \mathbf{C}[z_1, \ldots, z_d], \quad i = 1, \ldots, d,$$

where $\partial_i^k \bullet f = \frac{\partial^k f}{\partial z_i^k}$. When $n = 1$, a holonomic function is nothing but a solution of linear ordinary differential equation with polynomial coefficients. In this case, a local minimum can be obtained numerically by a difference scheme, which is called Euler's method. Readers may think that it will be straight forward to generalize Euler's method to d-variables, which we will call *holonomic gradient descent*. However, as we will see in this paper, a generalization of Euler's method to d-variables requires to utilize the theory, algorithms, and efficient implementations of Gröbner basis for holonomic systems, which have been studied recently (see [13] and its references).

In Section 2, we will illustrate holonomic gradient descent precisely. In Sections 3 and 4, we study the Fisher-Bingham integral as a holonomic function. The integral is important in the directional statistics. In Section 5, we consider problems in the directional statistics as applications of Sections 2, 3, and 4. In the last section 6, we will discuss advantages and disadvantages of our method.

2 Holonomic Gradient Descent

When we are given a Gröbner basis B, a set of monomials S is called the set of *standard monomials* of B if it is the set of the monomials which are irreducible (non-divisible) by B (see, e.g., [4], [15]). Let $g(z_1, \ldots, z_d)$ be a holonomic function and we suppose that it is annihilated by a holonomic ideal I. Let S be the set of standard monomials of a Gröbner basis of RI in R, which is a ring of differential operators with rational function coefficients. We may suppose that S contains 1 as the first element of S. Since the function g is holonomic, the column vector of functions $G = (s_i \bullet g \,|\, s_i \in S)^T$ satisfies the following set of linear partial differential equations (see, e.g., [13, p.39]).

$$\frac{\partial G}{\partial z_i} = P_i G, \quad i = 1, \ldots, d. \tag{2}$$

Note that each equation can be regarded as an ordinary differential equation with respect to z_i with parameters $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_d$. We call the system of differential equations (2) *the Pfaffian system* for g. The first entry of G, which is denoted by G_1, is g.

For holonomic functions, we can evaluate the gradient ∇g by normal forms with respect to a Gröbner basis. Let F_i be the normal form of ∂_i by a Gröbner

basis of RI in R. Then, F_i can be written as $\sum_{s_j \in S} a_{ij} s_j$ modulo RI where $a_{ij} \in \mathbf{C}(z_1, \ldots, z_d)$. Therefore, we have $\partial_i \bullet g = \sum_{s_j \in S} a_{ij}(s_j \bullet g)$, and then $\nabla g = (a_{ij})G$. This enables us to apply the standard gradient descent for holonomic functions as in the following algorithm.

Algorithm 1 Let $\varepsilon > 0$ be a small number.

1. Obtain a Gröbner basis of RI in R and the set of standard monomials S of the basis.
2. Compute the matrices P_i in (2) by the normal form algorithm and the Gröbner basis and the set of standard monomials.
3. Compute the normal form F_i of ∂_i by a Gröbner basis of RI in R and determine the matrix (a_{ij}). (Apply the normal form algorithm, which is also called the division algorithm or the reduction algorithm, for ∂_i in R. See, e.g., [15].)
4. Take a point c as a starting point and evaluate numerically G at $z = c$. Denote the value by $\bar{G}$ and put $e = c$.
5. Evaluate the approximate value $(a_{ij}(e))\bar{G}$ of the gradient $\tilde{g} = \nabla g$ at e. If $\tilde{g} = 0$, then stop.
6. Put $e \leftarrow e - \varepsilon\tilde{g}$ (move to $e - \varepsilon\tilde{g}$).
7. Obtain the approximate value $\bar{G}$ of G at $z = e$ by solving numerically ([10]) the Pfaffian system (2). Goto 5.

Holonomic functions are holomorphic out of the singular locus of the system of differential equations. We can apply known convergence criteria to this algorithm (see, e.g., [14]).

We suggest the second algorithm to find a local minimum (resp. maximum) of the holonomic function g in the domain $E = [a_1, b_1] \times \cdots \times [a_d, b_d]$. Although it is an analogous method and is less sophisticated than the Algorithm 1, it is faster in our applications discussed in the Section 5. We denote by εe_i a small vector $(0, \ldots, 0, \varepsilon, 0, \ldots, 0)$, $\varepsilon > 0$ in $\mathbf{R}^d$.

Algorithm 2

1. Obtain a Gröbner basis of RI in R and a set of standard monomials S of the basis.
2. Compute the matrices P_i in (2) by the normal form algorithm and the Gröbner basis and the set of standard monomials.
3. Take a point c in E as a starting point and evaluate numerically G at $z = c$. Denote the value by $\bar{G}$ and put $e = c$.

4. Compute the approximate value $\bar{G}_{\pm i}$ of G at $z = e \pm \varepsilon e_i$ by a difference scheme $\bar{G}_{\pm i} = S(\bar{G}(e), P_i(e), e, \pm\varepsilon)$. We evaluate it for all index i and the sings $+$ or $-$ as long as the point $e \pm \varepsilon e_i$ lies in the domain E.

5. Choose the index i and the sign $+$ or $-$ so that $(\bar{G}_{\pm i})_1$ is the minimum (resp. maximum) in $\{(\bar{G}_{\pm j})_1 \ (j = 1, \ldots, d), \bar{G}(e)_1\}$. If there is no such index, then stop and return the z value e and the first element $\bar{G}(e)_1$ of the vector $\bar{G}(e)$. If i and $\star$ are such index and sign, then put $e = e \star \varepsilon e_i$ and $G(e) = G_{\star i}$. Go to 4.

We denote by $z^{(i_1,\ldots,i_d)}$ the $(i_1, \ldots, i_d)$-th grid point in the domain E with the meshsize ε. In other words, we put $z^{(i_1,\ldots,i_d)} = (a_1, \ldots, a_d) + \sum_{k=1}^{d} i_k e_k \varepsilon$. Let $\bar{g}(z^{(i_1,\ldots,i_d)})$ be the value of g at the grid point $z^{(i_1,\ldots,i_d)}$ obtained by the difference scheme S. We are interested in the question whether e gives a local minimum point. The next theorem gives an answer to this question.

Theorem 1 *Suppose that the approximate values of g converge in the order more than 2. In other words, we have*

$$|\bar{g}(z^{(i_1,\ldots,i_d)}) - g(z^{(i_1,\ldots,i_d)})| \leq M\varepsilon^p, \quad p \geq 3 \tag{3}$$

uniformly on E. Let $z = e$ be the output grid point of the Algorithm 2. Then, there exists a non-negative constant δ such that the point e lies in the domain

$$\bigcap_{i=1}^{d} \left(\left\{ z \in E \;\middle|\; \left| \frac{\partial g}{\partial z_i}(z) \right| \leq \delta \right\} \cap \left\{ z \in E \;\middle|\; \frac{\partial^2 g}{\partial z_i^2}(z) \geq -\delta \right\} \right).$$

The constant δ converges to 0 when $\varepsilon \to 0$.

Proof. We denote by $g(z^{(i_1,\ldots,i_d)})$ the value of g at the grid point $z = z^{(i_1,\ldots,i_d)}$ and by $\bar{g}(z^{(i_1,\ldots,i_d)})$ the approximate value of g at the grid point $z = z^{(i_1,\ldots,i_d)}$ obtained by the difference scheme S. It follows from the assumption that we have the estimate

$$\begin{aligned} & \varepsilon^{-1}(g(e + \varepsilon e_i) - g(e)) \\ = \ & \varepsilon^{-1} \left((g(e + \varepsilon e_i) - \bar{g}(e + \varepsilon e_i)) + (\bar{g}(e + \varepsilon e_i) - \bar{g}(e)) + (\bar{g}(e) - g(e)) \right) \\ \geq \ & -2M\varepsilon^{p-1} \end{aligned}$$

and

$$\varepsilon^{-1}(g(e - \varepsilon e_i) - g(e)) \geq -2M\varepsilon^{p-1}.$$

Since $\varepsilon^{-1}(g(e + \varepsilon e_i) - g(e))$, which is $\geq -2M\varepsilon^{p-1}$, and $-\varepsilon^{-1}(g(e - \varepsilon e_i) - g(e))$, which is $\leq 2M\varepsilon^{p-1}$, converge to $\frac{\partial g}{\partial z_i}(e)$, we conclude that the value of the partial derivative of g at $z = e$ stays in a neighborhood of 0.

Let us proceed on an estimation of $\frac{\partial^2 g}{\partial^2 z_i}$, which is approximated by $\varepsilon^{-2}(g(e + \varepsilon e_i) - 2g(e) + g(e - \varepsilon e_i))$. This is estimated as follows.

$$\varepsilon^{-2}(g(e + \varepsilon e_i) - 2g(e) + g(e - \varepsilon e_i))$$

$$
\begin{aligned}
&= \varepsilon^{-2}(g(e+\varepsilon e_i) - g(e) + g(e-\varepsilon e_i) - g(e)) \\
&= \varepsilon^{-2}((g(e+\varepsilon e_i) - \bar{g}(e+\varepsilon e_i)) + (\bar{g}(e+\varepsilon e_i) - \bar{g}(e)) + (\bar{g}(e) - g(e)) \\
&\qquad +(g(e-\varepsilon e_i) - \bar{g}(e-\varepsilon e_i)) + (\bar{g}(e-\varepsilon e_i) - \bar{g}(e)) + (\bar{g}(e) - g(e))) \\
&\geq -4M\varepsilon^{p-2}
\end{aligned}
$$

Therefore, we can see that the second partial derivative of g at $z = e$ is bounded from below by a small constant $-\delta$. Q.E.D.

The assumption (3) is satisfied when we use the 4th order Runge-Kutta method and $P_i(z)$ are holomorphic on E. In this case, $p = 4$.

We note that the approximate minimal point $z = e$ does not always converge to a point. For example, suppose that $g(z_1, z_2) = (z_2 - z_1)^2$ and $E = [0, 1] \times [0, 1]$. Then, the minimum of this function is attained on $z_2 = z_1$ and the approximate minimal point $z = e$ will stay in a neighborhood of this line, but it might not converge to a point.

It is easy to generalize the algorithm for holonomic function which satisfies inhomogeneous holonomic system.

Remark 1 In our implementation, we do not evaluate new G for all directions. If the direction e_i is chosen, then we move to the direction as long as g decreases to the direction e_i. Because P_k is usually a matrix of a huge size and the computational cost of restricting the variables z_j, $j \neq i$ in P_k to numbers is extremely high. The standard gradient descent moves from the point p to the direction $-\nabla g$. Our method does not use this direction because it requires the evaluation of new G for all directions.

The holonomic gradient descent is Euler's method when the number of variables is 1. In applications, the function to minimize is often given as a definite integral with parameters, for which we can utilize algorithms for holonomic systems to find a differential equation. We revisit this example in the Algorithm 3.

Example 1 $d = 1$. $g(x) = \exp(-x+1)\int_0^\infty \exp(xt - t^3)dt$. The function $g(x)$ satisfies the differential equation $(3\partial_x^2 + 6\partial_x + (3-x)) \bullet g = \exp(-x+1)$, which can be obtained by an integration algorithm for D-modules [8]. The holonomic rank is 2 and we use a set of standard monomials $S = \{1, \partial_x\}$ and we have

$$
\frac{dG}{dx} = \begin{pmatrix} 0 & 1 \\ (-3+x)/3 & -2 \end{pmatrix} G + \begin{pmatrix} 0 \\ \exp(-x+1)/3 \end{pmatrix}
$$

This system is obtained by the normal form algorithm in the ring R [12]. We evaluate $G(0) = (g(0), g'(0))^T$ by a numerical integration method; $\bar{G}(0) = (2.427, -1.20)^T$. We apply the holonomic gradient descent in $D = [0, 5]$ with $\varepsilon = 0.1$ and the 4th order Runge-Kutta method and obtain $x = e = 3.4$ and $g(e) = 1.016$ as the minimum in this domain.

3 Fisher-Bingham Integral on S^n

We denote by $S^n(r)$ the n-dimensional sphere with the radius r in the $n+1$ dimensional Euclidean space. Let x be a $(n+1) \times (n+1)$ symmetric matrix and y a row vector of length $n+1$. We are interested in the following integral with the parameters x, y, r.

$$F(x, y, r) = \int_{S^n(r)} \exp(t^T x t + y t)|dt| \tag{4}$$

Here, t is the column vector $(t_1, \ldots, t_{n+1})^T$ and $|dt|$ is the standard measure on the sphere. For example, in case of $n = 1$, the measure $|dt|$ is $rd\theta$ in the polar coordinate system $t_1 = r\cos\theta, t_2 = r\sin\theta$. We call the integral (4) *the Fisher-Bingham integral* on the sphere $S^n(r)$.

We denote by x_{ii} the i-th diagonal entry of the matrix x and by $x_{ij}/2$ the (i, j)-th entry (or (j, i)-th entry) of the matrix x. Then, we can regard the function (the Fisher-Bingham integral) $F(x, y, r)$ as the function of x_{ij} $(1 \leq i \leq j \leq n+1)$ and y_i $(1 \leq i \leq n+1)$ and r.

Theorem 2 *The Fisher-Bingham integral $F(x, y, r)$ is a holonomic function.*

Proof. We will prove it for $n = 1$ to avoid complicated indices. The cases for $n > 1$ can be shown analogously.

Put $x_1 = r\cos\theta, x_2 = r\sin\theta$ (the polar coordinate system). Then, the invariant measure $|dt|$ is written as $rd\theta$. Therefore, $F(x, y, r) = \int_0^{2\pi} e^{g(x,y,r,\theta)} r d\theta$ where $g(x, y, r, \theta) = x_{11}r^2\cos^2\theta + x_{12}r^2\cos\theta\sin\theta + x_{22}r^2\sin^2\theta + y_1 r\cos\theta + y_2 r\sin\theta$. If we put $s = \tan\frac{\theta}{2}$, then $\sin\theta = 2s/(s^2+1)$ and $\cos\theta = (1-s^2)/(s^2+1)$ and $d\theta = \frac{2}{1+s^2}ds$ (rational representation of trigonometric functions). Then, the integral $F(x, y, r)$ can be written as

$$\int_{-\infty}^{\infty} h(x, y, r, s)ds, \quad h = e^{\tilde{g}(x,y,r,s)}\frac{2}{1+s^2}$$

where $\tilde{g}$ is a rational function in x, y, r, s. It is known that the exponential of a rational function is a holonomic function and the product of holonomic functions is a holonomic function, then the integrand is a holonomic function in x, y, r, s (see, e.g., [10] and [11]). By Lemma 1 in the Appendix, there exists a differential operator $\ell(x, y, r, \partial_{x_{ij}}) - \partial_s \ell_1(x, y, r, \partial_{x_{ij}}, \partial_s)$ depending only on $x, \partial_{x_{ij}}, y, r, \partial_s$ which annihilates the integrand h. Therefore, we have $\ell \bullet F(x, y, r) = [\ell_1 \bullet h]_{-\infty}^{\infty}$. Since we can show that $\partial_{x_{ij}}^m \partial_s^n \bullet h$ is a finite holonomic function at $s = \pm\infty$ for any non-negative integers m and n, the function $F(x, y, r)$ is annihilated by an ordinary differential operator of $\partial_{x_{ij}}$ with parameters x, y, r. The existence of annihilating ordinary differential operators with respect to ∂_{y_i} and ∂_r can be shown analogously. This existence implies that $F(x, y, r)$ is a holonomic function (see, e.g., [16, Theorem 2.4]). Q.E.D.

4 Holonomic system for the Fisher-Bingham Integral

In Example 1, we obtained a differential equation for the definite integral with parameters by a D-module algorithm. This algorithm works for any definite integral with a holonomic integrand, however, it requires huge computational resources. For the Fisher-Bingham integral, we can obtain a holonomic system of differential equations for the case of $n = 1$ by our computer program. The case of $n = 2$ is not feasible by our program. We obtain the following result for general n by utilizing an invariance of the Fisher-Bingham integral.

Theorem 3 *The function $F(x, y, r)$ is annihilated by the following system of linear partial differential operators.*

$$\partial_{x_{ij}} - \partial_{y_i}\partial_{y_j}, \quad (i \le j) \tag{5}$$

$$\sum_{i=1}^{n+1} \partial_{x_{ii}} - r^2, \tag{6}$$

$$x_{ij}\partial_{x_{ii}} + 2(x_{jj} - x_{ii})\partial_{x_{ij}} - x_{ij}\partial_{x_{jj}} + \sum_{k\neq i,j} (x_{jk}\partial_{x_{ik}} - x_{ik}\partial_{x_{jk}}) + y_j\partial_{y_i} - y_i\partial_{y_j}, \quad (i < j, x_{k\ell} = x_{\ell k}), \tag{7}$$

$$r\partial_r - 2\sum_{i\le j} x_{ij}\partial_{x_{ij}} - \sum_i y_i\partial_{y_i} - n. \tag{8}$$

We note that operators of the form (5) can be written as

$$\partial^u - \partial^v, \quad Au = Av,\ u, v \in \mathbf{N}^{(n+1)(n/2+2)}.$$

Here, A is the support matrix of the polynomial $t^T xt + yt$ with respect to t. For example, in case of $n = 1$, the polynomial is $x_{11}t_1^2 + x_{12}t_1t_2 + x_{22}t_2^2 + y_1t_1 + y_2t_2$ and the matrix A is

$$A = \begin{pmatrix} 2 & 1 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 & 1 \end{pmatrix}$$

of which column vectors stand for supports of the polynomial respectively.

Proof. Denote by $g(x, y, t) = \exp(t^T xt + yt)$ the integrand of (4). The operator $\partial_{x_{ij}} - \partial_{y_i}\partial_{y_j}$ annihilates $g(x, y, t)$ because $(\partial_{x_{ij}} - \partial_{y_i}\partial_{y_j}) \bullet g = (t_it_j - t_it_j)g = 0$. On the sphere $S^n(r)$, we have an identity $\sum_{i=1}^{n+1} t_i^2 = r^2$. Hence $\sum_{i=1}^{n+1} \partial_{x_{ii}} - r^2$ annihilates $g(x, y, t)$ for $t \in S^n(r)$.

Let us prove (7). By the invariance of the measure $|dt|$ with respect to the orthogonal group, we have $F(PxP^T, yP^T, r) = F(x, y, r)$ for any orthogonal transformation P on $S^n(r)$. Let I_{n+1} be the $(n+1)\times(n+1)$ identity matrix and e_{ij} be an $(n+1)\times(n+1)$ matrix whose (k, l)-th entry $(e_{ij})_{kl}$ is 1 if $(i, j) = (k, l)$ and 0 else. Put $P = \begin{pmatrix} \cos\epsilon & -\sin\epsilon \\ \sin\epsilon & \cos\epsilon \end{pmatrix} \oplus I_{n-1}$. This is an $(n + 1) \times (n + 1)$

orthogonal matrix and we have $P = I_{n+1} + \epsilon(e_{12} - e_{21}) + O(\epsilon^2)$. Hence we have

$$\begin{aligned} PxP^T &= (I + \epsilon(e_{12} - e_{21}))x(I + \epsilon(e_{21} - e_{12})) + O(\epsilon^2) \\ &= x + \epsilon(e_{12}x - e_{21}x + xe_{21} - xe_{12}) + O(\epsilon^2) \\ &= x + \epsilon \sum_{i \le j} f_{ij}(x)(e_{ij} + e_{ji})/2 + O(\epsilon^2), \end{aligned}$$

where

$$f_{ij}(x) = \begin{cases} x_{12} & \text{if} \quad i = j = 1, \\ 2(x_{22} - x_{11}) & \text{if} \quad i = 1, j = 2, \\ -x_{12} & \text{if} \quad i = j = 2, \\ x_{2j} & \text{if} \quad i = 1, j \ge 3, \\ -x_{1j} & \text{if} \quad i = 2, j \ge 3, \\ 0 & \text{if} \quad j \ge i \ge 3, \end{cases}$$

and

$$yP^T = y + \epsilon \begin{pmatrix} y_2 & -y_1 & 0 \end{pmatrix} + O(\epsilon^2).$$

Differentiating the identity $F(PxP^T, yP^T, r) - F(x, y, r) = 0$ by ϵ, we obtain

$$0 = \left(\sum_{i \le j} f_{ij}(x) \partial_{x_{ij}} + y_2 \partial_{y_1} - y_1 \partial_{y_2} \right) \bullet F + O(\epsilon).$$

Taking the limit $\epsilon \to 0$, we have (7) with $i = 1$ and $j = 2$. By symmetry we have (7) for any $i < j$.

Finally we differentiate the identity $\rho^n F(\rho^2 x, \rho y, r) = F(x, y, \rho r)$ by ρ and take the limit $\rho \to 1$. Then, we obtain

$$\left(n + 2 \sum_{i \le j} x_{ij} \partial_{x_{ij}} + \sum_i y_i \partial_{y_i} \right) \bullet F = r \partial_r \bullet F$$

This shows that F is annihilated by (8). Q.E.D.

Example 2 When $n = 1$, the system is written as follows.

$$\begin{aligned} &\partial_{x_{11}} - \partial_{y_1}^2, \partial_{x_{12}} - \partial_{y_1}\partial_{y_2}, \partial_{x_{22}} - \partial_{y_2}^2, \\ &\partial_{x_{11}} + \partial_{x_{22}} - r^2, \\ &x_{12}\partial_{x_{11}} + 2(x_{22} - x_{11})\partial_{x_{12}} - x_{12}\partial_{x_{22}} + y_2\partial_{y_1} - y_1\partial_{y_2}, \\ &r\partial_r - 2(x_{11}\partial_{x_{11}} + x_{12}\partial_{x_{12}} + x_{22}\partial_{x_{22}}) - (y_1\partial_{y_1} + y_2\partial_{y_2}) - 1. \end{aligned}$$

Example 3 When $n = 2$, the system is written as follows.

$$\begin{aligned} &\partial_{x_{11}} - \partial_{y_1}^2, \partial_{x_{12}} - \partial_{y_1}\partial_{y_2}, \partial_{x_{13}} - \partial_{y_1}\partial_{y_3}, \\ &\partial_{x_{22}} - \partial_{y_2}^2, \partial_{x_{23}} - \partial_{y_2}\partial_{y_3}, \partial_{x_{33}} - \partial_{y_3}^2, \\ &\partial_{x_{11}} + \partial_{x_{22}} + \partial_{x_{33}} - r^2, \end{aligned}$$

$$
\begin{aligned}
&x_{12}\partial_{x_{11}} + 2(x_{22} - x_{11})\partial_{x_{12}} - x_{12}\partial_{x_{22}} + x_{23}\partial_{x_{13}} - x_{13}\partial_{x_{23}} + y_2\partial_{y_1} - y_1\partial_{y_2},\\
&x_{13}\partial_{x_{11}} + 2(x_{33} - x_{11})\partial_{x_{13}} - x_{13}\partial_{x_{33}} + x_{23}\partial_{x_{12}} - x_{12}\partial_{x_{23}} + y_3\partial_{y_1} - y_1\partial_{y_3},\\
&x_{23}\partial_{x_{22}} + 2(x_{33} - x_{22})\partial_{x_{23}} - x_{23}\partial_{x_{33}} + x_{13}\partial_{x_{12}} - x_{12}\partial_{x_{13}} + y_3\partial_{y_2} - y_2\partial_{y_3},\\
&r\partial_r - 2(x_{11}\partial_{x_{11}} + x_{12}\partial_{x_{12}} + x_{13}\partial_{x_{13}} + x_{22}\partial_{x_{22}} + x_{23}\partial_{x_{23}} + x_{33}\partial_{x_{33}})\\
&\quad -(y_1\partial_{y_1} + y_2\partial_{y_2} + y_3\partial_{y_3}) - 2.
\end{aligned}
$$

Let R be the ring of differential operators with rational function coefficients.

Proposition 1 *1. The operators given in Theorem 3 generate a holonomic ideal in case of $n = 1$ and $n = 2$.*

2. The holonomic rank of the system for $n = 1$ is 4. A set of standard monomials in R is

$$1, \partial_{y_1}, \partial_{y_2}, \partial_r.$$

3. The holonomic rank of the system for $n = 2$ is 6. A set of standard monomials in R is

$$1, \partial_r, \partial_{y_3}, \partial_{y_2}, \partial_{y_1}, \partial_{x_{33}}.$$

The proposition can be shown by a calculation on a computer with applying algorithms for holonomic systems [18, `toc.html`], [13].

We conjecture that the system of operators given in Theorem 3 generates a holonomic ideal in D.

5 Computational Results

Let us apply the holonomic gradient descent to minimize a holonomic function

$$F(x, y, 1) \exp\left(-\sum_{1 \le i \le j \le n} S_{ij}x_{ij} - \sum_i S_i y_i\right) \tag{9}$$

with respect to x and y for given data $((S_{ij})_{i\le j}, (S_i))$. Here $F(x, y, 1)$ is the Fisher-Bingham integral (4) with $r = 1$.

First we describe the background in statistics. This paragraph can be skipped for the reader interested only in computational results. *The Fisher-Bingham family* on the sphere $S^n(1)$ is defined by the set of probability density functions

$$p(t|x, y) = F(x, y, 1)^{-1} \exp(t^\top xt + yt) \tag{10}$$

with respect to the standard measure $|dt|$ on $S^n(1)$. Since $\int_{S^n(1)} p(t|x, y)|dt| = 1$, the function $p(t|x, y)$ is actually a probability density function. We note that the parameter x has redundancy. In fact, for any real number c the density function $p(t|x + cI, y)$ is equal to $p(t|x, y)$, where I denotes the identity matrix. *A sample* refers to a set of points $\{t(1), \dots, t(N)\}$ on $S^n(1)$, where $N \ge 1$ is called the sample size. Assume that the sample is distributed according to $\prod_{\nu=1}^N p(t(\nu)|x, y)$ (independently identically distributed). To estimate

the unknown parameter (x, y) from the sample is a main problem in statistics. An established method is *the maximum likelihood method (MLE)* that maximizes a function $\prod_{\nu=1}^{N} p(t(\nu)|x, y)$ with respect to (x, y). The MLE is equivalent to minimize the function (9) with $S_{ij} = N^{-1} \sum_{\nu=1}^{N} t_i(\nu) t_j(\nu)$ and $S_i = N^{-1} \sum_{\nu=1}^{N} t_i(\nu)$. It is known that the logarithm of (9) is convex (see e.g. [2]) and therefore a local minimum at an interior point is actually the global minimum. Although gradient systems on probability families for optimization are considered by [7], difficulty of computing the integral F is not taken into account. See [6] for details on the Fisher-Bingham family and other probability families on the sphere. We test two examples, astronomical data and magnetism data. The astronomical data consist of the locations of 188 stars of magnitude brighter than or equal to 3.0. The data is available from the Bright Star Catalog (5th Revised Ed.) distributed from the Astronomical Data Center. The magnetism data is analyzed in [3] and [5].

The data and programs to test the following examples can be obtained from [18]. Please look up the instruction in the files `fb-demo-0.txt` and `fb-demo-1.txt`.

Astronomical data: We consider the problem to minimize

$$F(x, y, 1) \exp\left(- \sum_{1 \le i \le j \le 3} S_{ij} x_{ij} - \sum_{i} S_i y_i \right)$$

on

$$\begin{aligned} & (x_{11}, x_{12}, x_{13}, x_{22}, x_{23}, x_{33}, y_1, y_2, y_3) \\ \in\ & [-30, 10] \times [-30, 10] \times [-30, 10] \times [-30, 10] \times [-30, 20] \times [-30, -0.01] \\ & \times [-30, -0.01] \times [-30, -0.001] \times [-30, 10] \end{aligned}$$

where

$$\begin{aligned} & (S_{11}, S_{12}, S_{13}, S_{22}, S_{23}, S_{33}, S_1, S_2, S_3) \\ =\ & (0.3119, 0.0292, 0.0707, 0.3605, 0.0462, 0.3276, -0.0063, -0.0054, -0.0762). \end{aligned}$$

The result is that the minimum 11.68573121328159669 is taken at $x = \begin{pmatrix} -0.161 & 0.3377/2 & 1.1104/2 \\ 0.3377/2 & 0.2538 & 0.6424/2 \\ 1.1104/2 & 0.6424/2 & -0.0928 \end{pmatrix}$, $y = (\underline{-0.019}, \underline{-0.0162}, -0.2286)$ with the grid size 0.05 and the 4th order Runge-Kutta method (see Fig. 1), where the values near the border are underlined. A starting point is found by a quadratic approximation of $F(x, y, 1)$, which is exactly calculated from the moments of the uniform distribution on the sphere, and solving the optimization problem for the quadratic polynomial.

We pose the conditions $x_{33} \le -0.01$, $y_1 \le -0.01$ and $y_2 \le -0.001$, because the variety $x_{33} = y_1 = y_2 = 0$ lies in the singularity of the Pfaffian system. The optimal point is found near $x_{33} = y_1 = y_2 = 0$ and the point is in the interior of

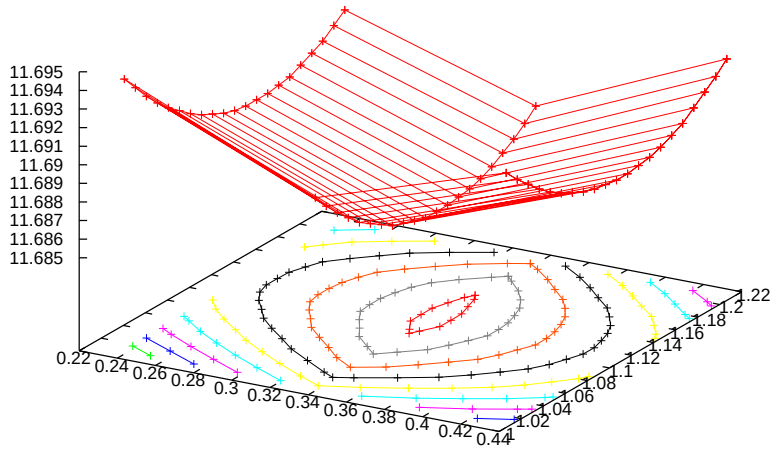

Figure 1: Graph of the target function with varying x_{12} and x_{13} around the minimal point for astronomical data.

the domain with the restriction $y_1 = -0.019$ and $y_2 = -0.0162$, which do not change from the starting values.

We briefly discuss the statistical meaning of the result. The spectral decomposition of x is $x = \sum_{i=1}^{3} \lambda_i z_i z_i^T$ with

$$(\lambda_1, \lambda_2, \lambda_3) = (0.7047, -0.0103, -0.6944)$$

and

$$(z_1, z_2, z_3) = \begin{pmatrix} -0.5063 & 0.5055 & 0.6987 \\ -0.6181 & -0.7777 & 0.1148 \\ -0.6014 & 0.3737 & -0.7061 \end{pmatrix}.$$

From the decomposition the density function (10) is high around $\pm z_1$ and low around $\pm z_3$. The effect of y is small because $|y| = 0.230$ is smaller than $|\lambda_i|$'s.

Magnetism data

We consider the problem to minimize

$$F(x, y, 1) \exp\left(- \sum_{1 \le i \le j \le 3} S_{ij} x_{ij} - \sum_i S_i y_i \right)$$

on

$$(x_{11}, x_{12}, x_{13}, x_{22}, x_{23}, x_{33}, y_1, y_2, y_3)$$

$$\in [-30,30]\times[-30,30]\times[-30,30]\times[-30,30]\times[-30,30]\times[-30,-0.01]$$
$$\times[-30,30]\times[-32,-0.001]\times[-30,32]$$

where

$$(S_{11},S_{12},S_{13},S_{22},S_{23},S_{33},S_1,S_2,S_3)$$
$$= (0.045,-0.075,0.014,0.921,-0.122,0.034,0.082,-0.959,0.131).$$

The result is that the minimum 0.4373096253840751950 is taken at
$x = x_o = \begin{pmatrix} 7.065 & -0.032/2 & 3.422/2 \\ -0.032/2 & 5.339 & 24.922/2 \\ 3.422/2 & 24.922/2 & -13.693 \end{pmatrix}$, $y = (1.642, \underline{-31.99}, \underline{31.992})$
with the grid size 0.01 and the 4th order Runge-Kutta method. Although y_2 and y_3 are on the border with this grid size, we can observe that the change of the target value is relatively small, when we enlarge the domain. In fact, we started the holonomic gradient descent from the optimal point, obtained by Wood's method [17], [18, `toc.html`], which is
$x = \begin{pmatrix} 5.985 & 8.478/2 & 2.902/2 \\ 8.478/2 & 6.869 & 16.732/2 \\ 2.902/2 & 16.732/2 & -12.853 \end{pmatrix}$, $y = (9.762, -28.770, 24.142)$. The optimal value of the target function is 0.4421940620633763292. If we restart the holonomic gradient descent from the point x_o by recalculating the integral values, we get a new optimal point and the target value changes only about 10^{-5}. Since the significant figures of the given data S_{ij}, S_i are 2 digits, we may conclude that there seems to be a variety which gives the optimal value of the target function. Our method finds a point in the variety and moves in the variety.

6 Comparison with Other Methods

The statistical problems considered in Section 5 can be solved by a different method. A. T. A. Wood [17] expressed the Fisher-Bingham integral of the case $n = 2$ as a single integral with the integrand expressed by a modified Bessel function. He gives a method to solve a minimization problem equivalent to our problem (9) based on this single integral representation. We implement his method by the statistical computing system R and obtain analogous computational results with us. The program is obtainable from [18, `toc.html`].

Although our two statistical problems can be solved by his different method, the advantage of our approach is that the holonomic gradient descent can be applied to a broad class of maximum likelihood problems. Let us illustrate our algorithm in the most general form.

Algorithm 3 (Holonomic gradient descent in the most automatic form)
Input: a definite integral $F(x) = \int_C f(x,t)dt$ with parameters $x = (x_1,\ldots,x_n)$ where $f(x,t)$ is a holonomic function of which annihilating ideal is J.
A holonomic function $g(x)$ of which annihilating ideal is J'.
Output: An approximate local minimum of $g(x)F(x)$ if the algorithm does not fail.

1. Apply integration algorithms for the holonomic ideal J (see, e.g., [1], [8], [9], [10], [13] and their references) to find a holonomic ideal $\int J$ annihilating the function $F(x)$. We note that these algorithms require some conditions for the domain of the integration C. If C does not satisfy these conditions, the algorithm fails.

2. Obtain a holonomic ideal I which annihilates $g(x)F(x)$ from $\int J$ and J' (see, e.g., [16], [10]).

3. Apply Algorithm 1 or Algorithm 2 for I where starting values of $F(x)$ and its derivatives are computed by a numerical integration method. If we have a numerical difficulty in these algorithms, the algorithm fails.

Example 1 illustrates this algorithm in the simplest form. The algorithm 3 works for a broad class of problems including the Fisher-Bingham integral, but we have several computational bottlenecks. For example, the step 1 (the integration algorithm) for the Fisher-Bingham integral can be easily performed in the case of $n = 1$ by our implementations, but the case $n = 2$ requires huge memory space and we could not finish the computation. In order to avoid this difficulty, we perform the step 1 of Algorithm 3 by hand as Section 4.

7 Appendix: Introduction to Holonomic Ideals

Although we want to suppose people with different disciplines as readers of this paper, the theory and algorithms for holonomic ideals are not very popular and there is no introductory text for these subjects. We will present an introductory overview on holonomic ideals and algorithms for them (see [13] and its references for proofs and original articles). This appendix is independent of the main text, but it will help to read the main text.

We denote by D the ring of differential operators with polynomial coefficients

$$D = \mathbf{C}\langle x_1, \ldots, x_d, \partial_1, \ldots, \partial_d \rangle,$$

which is also called the Weyl algebra. This is an associative non-commutative ring and x_i and ∂_j have the commuting relations

$$x_i x_j = x_j x_i, \partial_i \partial_j = \partial_j \partial_i, \partial_i x_j = x_j \partial_i + \delta_{ij}$$

where δ_{ij} is Kronecker's delta. Elements in D are often expressed by using the multi-index notation such as $x^\alpha \partial^\beta = \prod_{i=1}^d x_i^{\alpha_i} \prod_{i=1}^d \partial_i^{\beta_i}$. By utilizing the commuting relations, any element of D can be transformed into the normally ordered form $\sum_{(\alpha,\beta)\in E} c_{\alpha\beta} x^\alpha \partial^\beta$. For example, the normally ordered form of $\partial_1 x_1 \partial_1$ is $x_1 \partial_1^2 + \partial_1$. Elements of D acts for function $f(x_1, \ldots, x_d)$ by

$$x^\alpha \partial^\beta \bullet f = x^\alpha \frac{\partial^{|\beta|} f}{\partial x_1^{\beta_1} \cdots \partial x_d^{\beta_d}}$$

where we denote by $\bullet$ the action.

Let us introduce one more important ring R, which we call the ring of differential operators with rational function coefficients,

$$R = \mathbf{C}(x_1, \ldots, x_d)\langle \partial_1, \ldots, \partial_d \rangle$$

where we denote by $\mathbf{C}(x_1, \ldots, x_d)$ the field of rational functions in $x_1, \ldots, x_d$. This is also an associative non-commutative ring and the commuting relations are $\partial_i \partial_j = \partial_j \partial_i$ and $\partial_i a(x) = a(x)\partial_i + \frac{\partial a}{\partial x_i}$ for $a(x) \in \mathbf{C}(x_1, \ldots, x_d)$.

The theory of Gröbner basis (see, e.g., [4]) can be easily generalized in D and R as long as orders satisfy some conditions. Since we do not need consider general orders, we fix the order to the graded reverse lexicographic order $\prec$ among monomials ∂^β in the sequel. In case of $d = 2$, we have

$$1 \prec \partial_2 \prec \partial_1 \prec \partial_2^2 \prec \partial_1 \partial_2 \prec \partial_1^2 \prec \cdots.$$

Let us explain some facts about Gröbner bases in R, which are used in this paper. For $f \in R$, the leading term (the initial term) with respect to $\prec$ is denoted by $\mathrm{in}_\prec(f)$ and we regard this element as an element in $\mathbf{C}(x_1, \ldots, x_d)[\xi_1, \ldots, \xi_d]$ where ξ_i and x_j commute each other. For example, when $f = (x_1 + x_2)\partial_1^2 \partial_2 + (x_2^4 + 1)\partial_2$, we have $\mathrm{in}_\prec(f) = (x_1 + x_2)\xi_1^2 \xi_2$. We say that $a(x)\xi^\beta$ divides $b(x)\xi^{\beta'}$ when $\beta_i \leq \beta_i'$ for all i. We call the following algorithm *the normal form algorithm* (*the division algorithm*).

Algorithm 4
Input: f, $G = \{g_1, \ldots, g_m\}$
Output: r (normal form) and $q_1, \ldots, q_m$ such that $f = \sum_{i=1}^m q_i g_i + r$ in R, $f \succeq q_i g_i$, and $\mathrm{in}_\prec(g_i)$ does not divide any term of r for all i.

1. $r \leftarrow 0$, $q_i \leftarrow 0$.
2. Call `wNormalForm`(f, G). We suppose that the output is $r', q_1', \ldots, q_m'$.
3. $f \leftarrow r' - \mathrm{in}_\prec(r')$, $r \leftarrow r + \mathrm{in}_\prec(r')$, $q_i \leftarrow q_i + q_i'$. If $f = 0$, then return $r, q_1, \ldots, q_m$ else goto 2.

Algorithm 5 (`wNormalForm`(f, G))

1. $r \leftarrow f$, $q_i \leftarrow 0$
2. *If* there exists i such that $\mathrm{in}_\prec(g_i)$ divides $\mathrm{in}_\prec(r)$ *then*
   $r \leftarrow r - c(x)\partial^\beta g_i$ where $c(x)\partial^\beta$ is chosen so that $\mathrm{in}_\prec(r) - c(x)\xi^\beta \mathrm{in}_\prec(g_i) = 0$.
   $q_i \leftarrow q_i + c(x)\partial^\beta$.
   else return $r, q_1, \ldots, q_m$.
3. goto 2.

Example 4 We compute the formal form of $f = \partial_1\partial_2^3$ by $g_1 = \underline{\partial_1\partial_2} + 1$, $g_2 = \underline{2x_2\partial_2^2} - \partial_1 + 3\partial_2 + 2x_1$. Since we have

$$\partial_1\partial_2^3 - \partial_2^2 g_1 = -\partial_2^2$$
$$-\partial_2^2 + \frac{1}{2x_2} g_2 = \frac{1}{2x_2}(-\partial_1 + 3\partial_2 + 2x_1) =: f^*,$$

the normal form is f^* and $q_1 = \partial_2^2$ and $q_2 = -\frac{1}{2x_1}$. We note that the set $\{g_1, g_2\}$ is a system of differential operators for the Bessel function in 2 variables (see, e.g., [10]).

Let I be a left ideal in R. A finite set $G = \{g_1, \ldots, g_m\}$, $g_i \in R$ is called a Gröbner basis of I with respect to $\prec$ when $\langle \mathrm{in}_\prec(g_1), \ldots, \mathrm{in}_\prec(g_m)\rangle = \langle \mathrm{in}_\prec(f) \,|\, f \in I\rangle$. Here, $\langle h_1, \ldots, h_m\rangle$ is the set $\sum_{i=1}^m \mathbf{C}(x_1, \ldots, x_d)[\xi_1, \ldots, \xi_d]h_i$, which is the ideal generated by $h_1, \ldots, h_m$ in $\mathbf{C}(x_1, \ldots, x_d)[\xi_1, \ldots, \xi_d]$. Although, we use $\langle , \rangle$ to specify the generators of non-commutative rings D and R, we also use the notation $\langle h_1, \ldots, h_m\rangle$ to denote the ideal generated by $h_1, \ldots, h_m$ here. It might be a little confusing, but the meaning will be clear in the context. A Gröbner basis can be obtained by the Buchberger algorithm. The proof is analogous with the case of the ring of polynomials (see, e.g., [4, Chapter 2]).

Let G a Gröbner basis. The element ∂^β is called a standard monomial when none of $\mathrm{in}_\prec(g)$, $g \in G$ divides ξ^β. The normal form is a sum of standard monomials over $\mathbf{C}(x_1, \ldots, x_d)$.

Example 5 This is a continuation of the previous example. Put $g_3 = \underline{\partial_1^2} - 3\partial_1\partial_2 - 2x_1\partial_1 + 2x_2\partial_2 - 2$. Then, the set $\{g_1, g_2, g_3\}$ is a Gröbner basis of the left ideal in R generated by g_1 and g_2. The set of the standard monomials is $\{1, \partial_1, \partial_2\}$.

The output r of the normal form algorithm depends on what index i we choose in the step 2 in the algorithm `wNormalForm`.

Theorem 4 *If G is a Gröbner basis of I, then the normal form r is unique.*

Proof. Suppose that we have two different normal forms r_1 and r_2. Since we have $r_1 - r_2 \in I$, $\mathrm{in}_\prec(r_1 - r_2)$ is divisible by an $\mathrm{in}_\prec(g_i)$ by the definition of Gröbner basis. But it contradicts to that r_i is a sum of standard monomials over $\mathbf{C}(x_1, \ldots, x_d)$. Q.E.D.

When the number of the standard monomials is finite, the ideal I is called *a zero-dimensional ideal*. It follows from Theorem 4 that the number is equal to the dimension of R/I as the vector space over $\mathbf{C}(x_1, \ldots, x_d)$ (see, e.g., [4, Chapter 5]). It implies that the number of the standard monomials does not depend on Gröbner bases.

We call $c(x)\partial^\beta$, $0 \neq c(x) \in \mathbf{C}(x_1, \ldots, x_d)$, a non-monic standard monomial when ∂^β is a standard monomial. Let $S = \{s_1 = 1, s_2, \ldots, s_p\}$ be a set of (independent) non-monic standard monomials of the Gröbner basis G such that $p = \sharp S = \dim_{\mathbf{C}(x_1,\ldots,x_d)} R/RG$. Put $Q = (s_i \bullet g \,|\, s_i \in S)^T$. In order to apply

holonomic gradient descent, we need to compute the $p \times p$ matrix P_i in the Pfaffian equations

$$\frac{\partial Q}{\partial z_i} = P_i Q, \quad i = 1, \ldots, d.$$

which is (2) in the main text. To obtain the matrix P_i, we apply the normal form algorithm to $\partial_i s_j$. Then, the coefficient of the normal form of $\partial_i s_j$ with respect to s_k is the (j,k)-th element of P_i. This is the step 2 of the Algorithm 1 in the main text.

Example 6 This is a continuation of the previous example. We choose $S = \{1, x_1\partial_1, x_2\partial_2\}$. Then, we obtain

$$P_1 = \begin{pmatrix} 0 & \frac{1}{x} & 0 \\ -x & \frac{2x^2+1}{x} & -2x \\ -y & 0 & 0 \end{pmatrix}, P_2 = \begin{pmatrix} 0 & 0 & \frac{1}{y} \\ -x & 0 & 0 \\ -x & \frac{\frac{1}{2}}{x} & \frac{\frac{-1}{2}}{y} \end{pmatrix}$$

where $x = x_1$ and $y = x_2$. We can utilize several packages to perform this computation. Among them, we use the package "yang" [12] on `Risa/Asir`[1], because it can perform a large scale computation, which is required in our applications. The code to obtain the result above is

```
import("yang.rr");
def ex1() {
 yang.define_ring([x,y]);
 L1=dx*dy+1;
 L2=dx^2-2*x*dx+2*y*dy+1;
 L3=2*y*dy^2+3*dy-dx+2*x;
 L=[L1,L2,L3];
 L=yang.util_pd_to_euler(L,[x,y]);
 L=map(nm,L);
 L=map(dp_ptod,L,[dx,dy]);
 G=yang.buchberger(L);
 S1=yang.constant(1);
 Sx=yang.operator(x);
 Sy=yang.operator(y);
 Base=[S1,Sx,Sy];
 Pf=yang.pfaffian(Base,G);
 return Pf;
}
ex1();
```

We need no application of the normal form algorithm for the step 3 of Algorithm 1 in this example. In fact, we have $\partial_1 = \frac{1}{x_1} s_2$ and $\partial_2 = \frac{1}{x_2} s_2$. Then, the matrix (a_{ij}) is $\begin{pmatrix} 0 & \frac{1}{x_1} & 0 \\ 0 & 0 & \frac{1}{x_2} \end{pmatrix}$.

We call a function F *a holonomic function* when it satisfies ordinary differential equations for all variables. In other words, F satisfies

$$\sum_{k=0}^{r_i} a_k^i(x_1, \ldots, x_d)\partial_i^k \bullet F = 0, \quad a_k^i \in \mathbf{C}[x_1, \ldots, x_d], \quad i = 1, \ldots, d. \tag{11}$$

[1] `http://www.openxm.org`

The set of operators in R which annihilate a function F is a left ideal in R. In fact, if $\ell_1 \bullet F = \ell_2 \bullet F = 0$, then we have $(\ell_1 + \ell_2) \bullet F = 0$, and if $\ell \bullet F = 0$, then $(h\ell) \bullet F = 0$ for all $h \in R$. We denote the set by $\mathrm{Ann}_R F$. When the function F is holonomic, $\mathrm{Ann}_R F$ contains ordinary differential equations (11). Therefore, the number of standard monomials of a Gröbner basis of $\mathrm{Ann}_R F$ is less than or equal to $\prod_{i=1}^d r_i$. In other words, we have $\dim_{\mathbf{C}(x_1,\ldots,x_d)} R/\mathrm{Ann}_R F \leq \prod_{i=1}^d r_i$. Conversely, we have the following theorem.

Theorem 5 *Let I be a left ideal in R. If $m := \dim_{\mathbf{C}(x_1,\ldots,x_d)} R/I$ is finite, then the left ideal I contains ordinary differential operators for all variables.*

Proof. $1, \partial_i, \partial_i^2, \ldots, \partial_i^m$ are linearly dependent in R/I, which we regard as a vector space over $\mathbf{C}(x_1, \ldots, x_d)$. This implies that there exist rational functions $c_k(x)$ such that $\sum_{k=0}^m c_k(x)\partial_i^k \in I$. Q.E.D.

This theorem is an analogy of the elimination theorem. The elimination in R can be done by an analogous method in case of the ring of polynomials (see, e.g., [4, Chapter 3]).

We have worked in the ring R. If we need to consider integrals of F, we need the theory and algorithms for the Weyl algebra D. Let us proceed on a discussion on D.

We first note that we can easily generalize the Gröbner basis theory for term orders $\prec$ in D. For example, in case of $d = 2$, the Gröbner basis theory works for the graded reverse lexicographic order such that $1 \prec x_1 \prec x_2 \prec \partial_1 \prec \partial_2 \prec x_1^2 \prec \cdots$. We note that for advanced algorithms like integration algorithms ([9], [8])) in D non-term orders are needed (see, e.g., [13, Chapter 1]).

We introduce the notion of a holonomic ideal. Let F_k be the set of elements in D of which order is less than or equal to k. In other words, F_k is a $\mathbf{C}$-vector space spanned by $x^\alpha \partial^\beta$, $|\alpha| + |\beta| \leq k$. $\{F_k\}$ is called the Bernstein filtration. A left ideal I in D is called *a holonomic ideal* when $\dim_{\mathbf{C}} F_k/F_k \cap I = O(k^d)$ for sufficiently large numbers k. The quotient D/I is called *a holonomic D-module* when I is a holonomic ideal. We note that the dimension agrees with the number of standard monomials of total degree less than or equal to k with respect to a Gröbner basis of I by the graded reverse lexicographic order (see, e.g., [4, Chapter 9]).

Lemma 1 *Let I be a holonomic ideal in the ring of differential operators $D = \mathbf{C}\langle x_1, \ldots, x_d, \partial_1, \ldots, \partial_d\rangle$. We choose a set of $d+1$ variables from the set $\{x_1, \ldots, x_d, \partial_1, \ldots, \partial_d\}$ and denote it by V. Then, the elimination ideal $I \cap \mathbf{C}\langle V\rangle$ contains a non-zero element.*

Proof. Consider the $\mathbf{C}$-linear map

$$\rho_k : \mathbf{C}\langle V\rangle \cap F_k \ni \ell \mapsto [\ell] \in F_k/F_k \cap I$$

The dimension of the $\mathbf{C}$-vector space $\mathbf{C}\langle V\rangle \cap F_k$ is $\binom{d+1+k}{d+1} = O(k^{d+1})$. On the other hand, we have $\dim_{\mathbf{C}} F_k/F_k \cap I = O(k^d)$ because I is a holonomic ideal. Since $\dim_{\mathbf{C}} \mathrm{Im}\, \rho_k = \dim_{\mathbf{C}} \mathbf{C}\langle V\rangle \cap F_k - \dim_{\mathbf{C}} \mathrm{Ker}\, \rho_k$, we conclude that the vector space $\mathrm{Ker}\, \rho_k$ contains a non-zero element. Q.E.D.

When I is a holonomic ideal, the number of standard monomials is infinite in general. It is natural to ask if there is a zero-dimensional ideal in D. However, the following theorem claims that the holonomic ideals are biggest ideals and there is no zero-dimensional ideal in D

Theorem 6 (Bernstein inequality) *Let I be a left ideal in D. Suppose that $I \neq D$. There exists a constant p such that $\dim_{\mathbf{C}} F_k/F_k \cap I = O(k^p)$ for sufficiently large k and the inequality $p \geq d$ holds.*

Let us explain a relation of a holonomic ideal in D and the zero dimensional ideal in R. For a left ideal I in D, put $RI = \{\sum_{\text{finite sum}} r_i f_i \,|\, r_i \in R, f_i \in I\}$. This is a left ideal in R. It follows from the Lemma 1 that if I is a holonomic ideal, then I contains ordinary differential operators for all variables and RI is a zero-dimensional ideal. Conversely, we have the following theorem.

Theorem 7 *If J is a zero-dimensional ideal in R, then $J \cap D$ is a holonomic ideal in D.*

An elementary proof of this fact is found in the appendix of [16]. We emphasize that when we are given a set of generators of J, it is not easy to find generators of $J \cap D$. The ideal $J \cap D$ is called *the Weyl closure* of J. An algorithm to construct this closure is given by H. Tsai (Algorithms for associated primes, Weyl closure, and local cohomology of D-modules. Lecture Notes in Pure and Appl. Math., 226, 169–194, Dekker, New York, 2002). For algorithms in D, we often require that *inputs are holonomic*. However, even finding a holonomic subideal of $J \cap D$ requires a high complexity. It often makes computational bottlenecks.

Example 7 We consider the function $f(x, y, z) = \exp(1/g)$ where $g = x^3 - y^2z^2$. The function f is annihilated by first order operators

$$g^2\partial_x + 3x^2, g^2\partial_y - 2yz^2, g^2\partial_z - 2y^2z$$

The left ideal I generated by these operator is not holonomic. The Weyl closure $J = RI \cap D$ is holonomic. The below is a `Macaulay 2`[2] script to check the holonomicity and find the Weyl closure of RI.

```
loadPackage "Dmodules"
D=QQ[x,y,z,dx,dy,dz, WeylAlgebra=>{x=>dx,y=>dy,z=>dz}];
I = ideal((x^3-y^2*z^2)^2*dx+3*x^2,
          (x^3-y^2*z^2)^2*dy-2*y*z^2,
          (x^3-y^2*z^2)^2*dz-2*y^2*z);
II=inw(I,{0,0,0,1,1,1});
print(dim II);  --- the output 4 implies that it is not holonomic.
J=WeylClosure I;
print(toString(J));
JJ=inw(J,{0,0,0,1,1,1});
print(dim JJ);  --- the output 3 implies that it is holonomic.
```

[2] `http://www.math.uiuc.edu/Macaulay2`

We close this appendix with introducing the integration ideal. The next fact is the fundamental fact for holonomic ideals and integrations.

Theorem 8 *If I is a holonomic ideal, then the integration ideal $(I+\partial_d D)\cap D_{d-1}$ is a holonomic ideal in D_{d-1}. Here $D_{d-1} = \mathbf{C}\langle x_1, \ldots, x_{d-1}, \partial_1, \ldots, \partial_{d-1}\rangle$.*

This theorem follows from the fact "if D/I is a holonomic D-module, then $D/(I + \partial_d D)$ is a holonomic D_{d-1} module". As to a proof of this fact, see, e.g., the Chapter 1 of the book "J. E. Björk, *Rings of Differential Operators*. North-Holland, New York, 1979".

Oaku's algorithm [9] to find integration ideals is explained in the Chapter 5 of [13] in a form relevant to our applications. Modifications of this algorithm [8] is used in the step 1 of our Algorithm 3 in the main text.

Example 8 Put $f(x,t) = \exp(xt - t^3)$. The function f is annihilated by the operators $\partial_t - (x - 3t^2)$, $\partial_x - t$, which generate a holonomic ideal L. This is a `Risa/Asir` code to find the integration ideal $(L + \partial_t \mathbf{C}\langle x, t, \partial_x, \partial_t\rangle) \cap \mathbf{C}\langle x, \partial_x\rangle$.

```
import("nk_restriction.rr");
def step1() {
 L=[dt-(x-3*t^2),
   dx-t];
 I=nk_restriction.integration_ideal(L,[t,x],[dt,dx],[1,0] |  inhomo=1);
 return I;
}
step1();
```

We write this introductory section with a few overlaps with [13]. For other fundamental facts, please refer to [13] and its references.

References

[1] M. Apagodu, D. Zeilberger, Multi-variable Zeilberger and Almkvist-Zeilberger algorithms and the sharpening of Wilf-Zeilberger theory, *Advances in Applied Mathematics* **37** (2006), 139-152.

[2] O. Barndorff-Nielsen, *Information and Exponential Families*, 1978, John Wiley & Sons.

[3] K. M. Creer, E. Irving, A. E. M. Nairn, Paleomagnetism of the great whin sill, *Geophysical Journal of the Royal Astronomical Society* **2** (1959), 306–323.

[4] D. Cox, J. Little, D. O'Shea, *Ideals, Varieties, and Algorithms*, Third Edition, 2007, Springer.

[5] J. T. Kent, The Fisher-Bingham Distribution on the Sphere, *Journal of the Royal Statistical Society. Series B* **44** (1982), 71–80.

[6] K. V. Mardia, P. E. Jupp, *Directional Statistics*, 2000, John Wiley & Sons.

[7] Y. Nakamura, Gradient systems associated with probability distributions, *Japan Journal of Industrial and Applied Mathematics* **11** (1994), 21–30.

[8] H. Nakayama, K. Nishiyama, An algorithm of computing inhomogeneous differential equations for definite integrals, `arXiv:1005.3417`

[9] T. Oaku, Algorithms for b-functions, restrictions, and algebraic local cohomology groups of D-modules, *Advances in Applied Mathematics* **19** (1997), 61–105.

[10] T. Oaku, Y. Shiraki, N. Takayama, Algebraic Algorithms for D-modules and Numerical Analysis, Z.M.Li, W.Sit (editors), Computer Mathematics, World scientific, 2003 (Proceedings of the sixth Asian symposium), 23–39,

[11] T. Oaku, N. Takayama, W. Walther, A localization algorithm for D-modules. *Journal of Symbolic Computation* **29** (2000) 721–728.

[12] K. Ohara, the yang package. `http://www.openxm.org`, `yang.rr`.

[13] M. Saito, B. Sturmfels, N. Takayama, *Gröbner Deformations of Hypergeometric Differential Equations*, 2000, Springer.

[14] A. Snyman, *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*, 2005, Springer.

[15] N. Takayama, Gröbner basis and the problem of contiguous relation, *Japan Journal of Applied Mathematics* **6** (1989), 147–160.

[16] N. Takayama, An approach to the zero recognition problem by Buchberger algorithm. *Journal of Symbolic Computation* **14** (1992) 265–282.

[17] A. T. A. Wood, Some notes on the Fisher-Bingham family on the sphere, *Communications in Statistics, Theory and Methods* **17** (1988), 3881–3897.

[18] `http://www.math.kobe-u.ac.jp/OpenXM/Math/Fisher-Bingham`.

Tomonari Sei, Akimichi Takemura ♯
Department of Mathematical Informatics
Graduate School of Information Science and Technolgy, University of Tokyo
Bunkyo, Tokyo, 113-0033, Japan

Nobuki Takayama[3♯] (`takayama@math.kobe-u.ac.jp`),
Hiromasa Nakayama♯, Kenta Nishiyama♯, Masayuki Noro♯[4]
Department of Mathematics, Kobe University
Rokko, Kobe, 657-8501, Japan

Katsuyoshi Ohara
Faculty of Mathematics and Physics, Kanazawa University
Kakuma-machi, Kanazawa, 920-1192, Japan

[3]Supported by Kakenhi 19204008
[4]Authors with ♯ belong to the JST crest Hibi project.