Holonomic methods in optimization, statistics, and machine learning

Nobuki Takayama (Department of Mathematics, Kobe University)

1. What is "holonomic"?
2. Holonomic gradient method (HGM) and maximal likelihood estimation (MLE) as a dynamical system
3. Some applications to optimization, statistics, and machine learning

Dimension in algebraic geometry

I : an ideal of $K[x] := K[x_1, \ldots, x_n]$. $K = \mathbb{C}$
$\mathrm{ord}_u x^\alpha := \langle u, \alpha \rangle$, $x^\alpha := \prod_{i=1}^n x_i^{\alpha_i}$
F_k : $\bigoplus_{\mathrm{ord}_{\mathbf{1}}(x^\alpha) \leq k} Kx^\alpha$, $\mathbf{1} := (1, \ldots, 1)$

Hilbert polynomial for I: $H(k) = \dim_K \frac{F_k}{F_k \cap I}$.
Example: $n = 2$, $I = \langle x_1 x_2 \rangle$.
$F_k/(F_k \cap I) = K + Kx_1 + \cdots + Kx_1^k + Kx_2 + \cdots + Kx_2^k$ then
$H(k) = 2k + 1$.

The degree of $H(k)$ w.r.t k is called the dimension of I. When $V(I)$ is a complex manifold, it agrees with the dimension as the complex manifold.

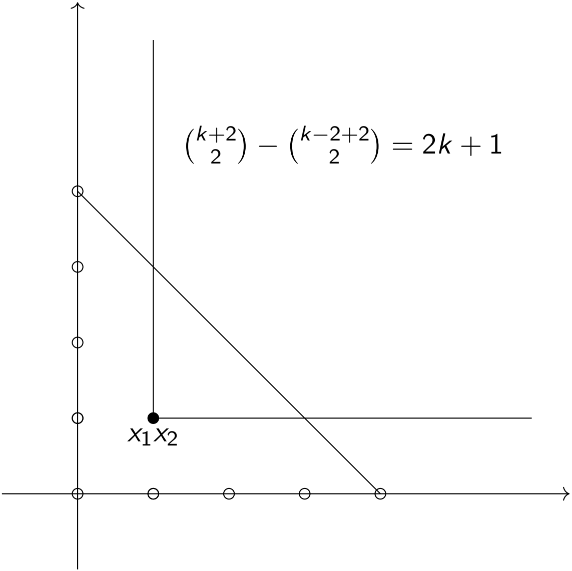


Figure: $\frac{F_4}{F_4 \cap I}$, $I = \langle x_1x_2 \rangle$

Weyl algebra (the ring of differential operators with polynomial coefficients)

$n = 1$, $x = x_1$.
$D_1 = K\langle x, \partial_x \rangle$, $\partial_x x = x\partial_x + 1$.
I : a left ideal of D_1. (1) $L, M \in I$, then $L - M \in I$. (2) If $L \in I$, then $ML \in I$ for any $M \in D_1$.
Any left ideal I of D_1 is generated by a finite set of operators $L_1, \ldots, L_m$. Gröbner basis method works.
f : a function. Action of ∂_x and x to f are defined respectively as $\partial_x \bullet f = \frac{df}{dx}$ and $x \bullet f = xf$.
If a function f is annihilated by $L_1, \ldots, L_m$, then f is annihilated by any element of I.
Example: $I = \langle x\partial_x - 1, \partial_x^2 \rangle$. $f = x$ is annihilated by I.

Any left ideal I ($\neq 0, D_1$) of D_1 is called holonomic ideal.

How to define "holonomic" ideal in several variable case?

Weyl algebra: $D_n = K\langle x_1, \ldots, x_n, \partial_1, \ldots, \partial_n \rangle$ where $x_i x_j = x_j x_i$, $\partial_i \partial_j = \partial_j \partial_i$, $\partial_i x_j = x_j \partial_i + \delta_{ij}$ $(\partial_i = \frac{\partial}{\partial x_i})$
$u, v \in \mathbb{R}^n$, $\mathrm{ord}_{(u,v)}(x^\alpha \partial^\beta) := \langle u, \alpha \rangle + \langle v, \beta \rangle$.
Definition (I.N.Bernstein=J.Bernstein, 1972).
$F_k = \bigoplus_{\mathrm{ord}_{(1,1)}(x^\alpha \partial^\beta) \leq k} K x^\alpha \partial^\beta$. Let I be a left ideal of D_n.
$H(k) = \dim_K \frac{F_k}{F_k \cap I}$. If the degree of the polynomial $H(k)$ is n, we call I a holonomic ideal.

Fact: Holonomic ideal I contains ordinary differential operators for all directions of the form

$$\sum_{k=0}^{r_i} s_{ik}(x) \partial_i^k, \quad s_{ik}(x) \in K[x],\ i = 1, \ldots, n \qquad (1)$$

Proof. K-linear map

$$p : F_k \cap K\langle x_1, \ldots, x_n, \partial_i \rangle \longrightarrow \frac{F_k}{F_k \cap I}$$

The dimension of the left hand side is $\binom{k+(n+1)}{n+1} = O(k^{n+1})$. Therefore, $p^{-1}([0])$ contains non-zero element. $\square$

An ideal generated by ordinary differential operators may not be holonomic. Example: $n = 2$.

$I = \langle L_1 := (x_1^3 - x_2^2)\partial_1 + 3x_1^2, L_2 := (x_1^3 - x_2^2)\partial_2 - 2x_2 \rangle$ $(I \bullet (x_1^3 - x_2^2)^{-1} = 0)$. $H(k) = \frac{k^3}{2} + 2k^2 + \frac{k}{2} + 2$. Then, I is not a holonomic ideal. Add $2x_1\partial_1 + 3x_2\partial_2 + 6$ to I, then it is a holonomic ideal.

Let R_n be the ring of differential operators with rational function coefficients; Rational Weyl algebra $R_n = \mathbb{C}(x)\langle \partial_1, \ldots, \partial_n \rangle$. J: generated by operators of the form (1).

Fact: $R_n J \cap D_n$ is holonomic ideal.

Example: $\frac{2}{p}x_1 L_1 + \frac{3}{p}x_2 L_2 = 2x_1\partial_1 + 3x_2\partial_2 + 6$, $p = x_1^3 - y_2^2$.

Holonomic ideal is a nice class

Theorem 1 (I.N.Bernstein, 1972[1])

1. *The degree of the Hilbert polynomial of a left ideal $I \subsetneq D_n$ of D_n is equal to n or more than n.*
2. *If I is holonomic in D_n, then $(I + x_n D_n) \cap D_{n-1}$ (restriction ideal) and $(I + \partial_n D_n) \cap D_{n-1}$ (integration ideal) are holonomic in D_{n-1}.*

Fact: If a rapidly decaying function f is annihilated by a holonomic ideal $I \subset D_n$, then the $n-1$ variables x' function $g(x') := \int_{-\infty}^{\infty} f(x) dx_n$ is annihilated by the integration ideal.

Proof. $L = L_1 + \partial_n L_2 \in (I + \partial_n D_n) \cap D_{n-1}$. Then

$$\begin{aligned} L \bullet g(x') &= \int_{-\infty}^{\infty} L_1 \bullet f \, dx_n + \int_{-\infty}^{\infty} \partial_n L_2 \bullet f \, dx_n = \int_{-\infty}^{\infty} \partial_n (L_2 \bullet f) \, dx_n \\ &= [L_2 \bullet f]_{-\infty}^{\infty} = 0 \qquad \square \end{aligned}$$

[1] Analytic continuation of generalized functions with respect to a parameter, Functional Analysis and Applications 6, 26-40

History

1. Mikio Sato: founder of algebraic anallysis. 1960's — 1990's.
2. M.Kashiwara, T.Kawai, J.Bernstein, Z.Mebkout, ... : the theory of D-modules, regular holonomic systems, ... Applications to algebraic geometry, representation theory, ... 1970's — the present.

---

3. D.Zeilberger, ... : holonomic method to prove and derive identities. 1990's — the present.
4. T.Oaku, N.T, U.Walther, ... : computational D-module theory. 1990's–2000's.
5. T.Sei, A.Takemura, T.Koyama, N.T., ... : holonomic gradient method (HGM) and holonomic gradient descent, 2010's — the present.

Holonomic distribution

Definition: Let f be a distribution. If f is annihilated by a holonomic ideal, then f is called a holonomic distribution. If f is a classical function, f is called a holonomic function. Roughly speaking[2],

1. A definite integral of a holonomic distribution is a holonomic distribution.
2. The sum, product (if it can be defined), derivatives of holonomic distributions are holonomic.

Example: $n = 1$, $x = x_1$. $Y(x) = 1$ $(x \geq 0)$, $Y(x) = 0$ $(x < 0)$ be the Heaviside function. $x\partial_x \bullet Y = 0$. Then $(x\partial_x - 1) \bullet xY(x) = 0$. The function $\sigma(x) = xY(x)$ is called ReLU (rectified linear unit) in machine learning.
$g(a, b, c) = \int_{\mathbb{R}^2} \exp(-au^2 - 2buv - cv^2)\sigma(u)\sigma(v)dudv$ is a holonomic function w.r.t. a, b, c.

[2] e.g., https://www.math.kobe-u.ac.jp/HOME/taka/ascm2003-paper.pdf

Exercise: Which are holonomic distributions?

1. $\exp(f(x_1, \ldots, x_n))$ where f is a rational function,
2. $\sin(x)$.
3. $\exp(x_1 \cos(t) + x_2 \sin(t))$
4. $\frac{1}{\sin x}$ [Hint] Use Th: Any solution of the ordinary differential equation $(a_m(x)\partial^m + \cdots + a_0(x)) \bullet f = 0$, $a_i \in \mathbb{C}[x]$, is holomorphic out of the singular locus $\{x \mid a_m(x) = 0\}$.
5. $\frac{1}{1+\exp(-x)}$ (sigmoid function).
6. $\Gamma(x)$, [Hint] $\Gamma(x)$ has poles at $x = -n$, $n \in \mathrm{N}_0$.
7. x^a where a is a constant,
8. $|x|$,
9. $\int_{-\infty}^{+\infty} \exp(-xt^6 - t)dt$, $x > 0$.

Todo, function graph.

2. Holonomic gradient method (HGM) and maximal likelihood estimation (MLE) as a dynamical system

Let I be a holonomic ideal in D_n which annihilates a holonomic function f. Then $R_n I$ is a zero dimensional ideal in R_n.
$r := \dim_{K(x)} \frac{R_n}{R_n I}$ is called a holonomic rank. $s_1, \ldots, s_r$: basis of $\frac{R_n}{R_n I}$. Put $F = (s_1 \bullet f, \ldots, s_r \bullet f)^T$. F satisfies Pfaffian equations:

$$\frac{\partial F}{\partial x_i} = P_i(x) F \qquad (2)$$

where P_i is a $r \times r$ matrix with rational function entries [3].

HGM to evalute $Z(x') = \int_D f(x) dx_{m+1} \cdots dx_n$ for holonomic f

(1) Compute the integration ideal (2) Derive Pfaffian equations. (3a) Evaluate a value of F at a relevant point. (3b) Extend the ~~value by numerically solving the~~ Pfaffian equations.

[3] P_i can be constructed by Gröbner basis.

Maximal likelihood estimation (MLE)

Example: unnormalized Von-Mises distribution on $S^1 \ni x$:
$u(\theta, x) = \exp(\theta_1 \cos x + \theta_2 \sin x)$.
(Holonomic) normalizing constant is

$$Z(\theta) = \int_0^{2\pi} \exp(\theta_1 \cos x + \theta_2 \sin x) dx$$

$F = (Z, \partial_1 Z)^T$, $\partial_i = \partial/\partial\theta_i$. Pfaffian system is

$$\begin{aligned} \frac{\partial F}{\partial \theta_1} &= \begin{pmatrix} 0 & 1 \\ \frac{\theta_1^2}{\theta_1^2+\theta_2^2} & \frac{\theta_2^2-\theta_1^2}{\theta_1(\theta_1^2+\theta_2^2)} \end{pmatrix} F =: P_1 F \\ \frac{\partial F}{\partial \theta_2} &= \begin{pmatrix} 0 & \theta_2/\theta_1 \\ \frac{\theta_1\theta_2}{\theta_1^2+\theta_2^2} & \frac{-2\theta_2}{\theta_1^2+\theta_2^2} \end{pmatrix} F =: P_2 F \end{aligned}$$

Fisher's MLE. X_i is observed data. Find θ which maximizes the likelihood

$$\ell(\theta; X) = \prod_{i=1}^{N} \frac{u(\theta, X_i)}{z(\theta)}$$

Let $f = \log \ell(\theta; X)$ be the log likelihood. The gradient descent updates $\theta = (\theta_1, \theta_2)$ by (new θ) $= \theta + \alpha \nabla_\theta f$. Then[4],

$$\dot{\theta} = \nabla_\theta f = \sum_{i=1}^{n} \frac{\nabla_\theta u}{u} - n \frac{\nabla_\theta F_1}{F_1}$$

From the chain rule and the Pfaffian equations,

$$\dot{F}_i = \dot{\theta}_1 (P_1 F)_i + \dot{\theta}_2 (P_2 F)_i$$

$$\begin{aligned}
\dot{\theta}_1 &= \; = \sum_{i=1}^{N} \cos(X_i) - N \frac{(P_1 F)_1}{F_1} && (3) \\
\dot{\theta}_2 &= \; = \sum_{i=1}^{N} \sin(X_i) - N \frac{(P_2 F)_1}{F_1} && (4) \\
\dot{F}_i &= \left(\sum_{i=1}^{N} \cos(X_i) - N \frac{(P_1 F)_1}{F_1} \right) (P_1 F)_i + \left(\sum_{i=1}^{N} \sin(X_i) - N \frac{(P_2 F)_1}{F_1} \right) (P_2 F)_i && (5)
\end{aligned}$$

$i = 1, 2.$

[4] U.Helmke, J.Moore, Optimization and Dynamical Systems, 1994

MLE for Von-Mises distribution

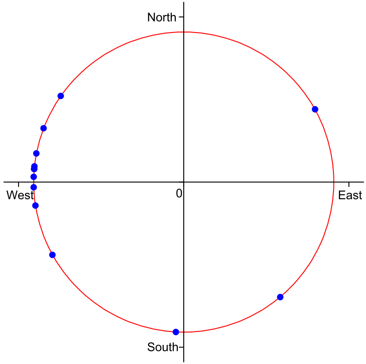


Figure: Wind direction at 10,000 meters above Sapporo, AM 9, 2011/1/1–2011/1/14 (1/11 missing)

$\mathrm{maxarg}_\theta \prod_{i=1}^{13} \frac{u(\theta, X_i)}{z(\theta)}$ where X_i is the direction in the figure.

Vector field of (3), (4), (5)

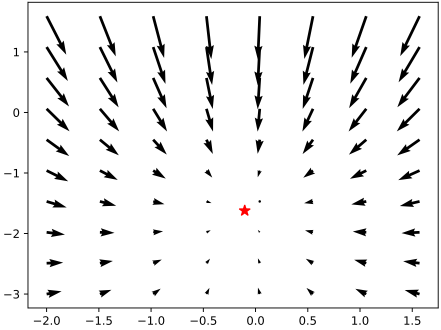


Figure: Vector field on (θ_1, θ_2) space

Solving (3), (4), (5),
by the initial value $(\theta; F) = (-1.62, -0.1; 9.82246, -6.12855)$, we have $\theta = (\theta_1, \theta_2) = (-0.1038, -1.6228)$.

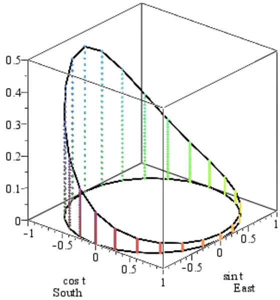


Figure: Wind direction distribution estimated by MLE

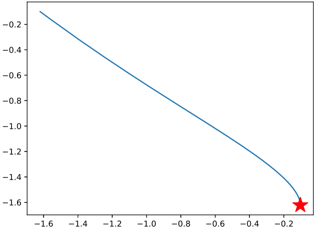


Figure: $\theta(t)$

Maximal likelihood estimation (MLE) as a dynamical system

Theorem 2

If an unnormalized distribution $u(\theta, x)$ is a holonomic distribution[5]*, then MLE w.r.t $u(\theta, x)$ and data in x space can be described by a dynamical system.*

[5] and $\int_\Omega u(\theta, x)dx$ satisfies the integration ideal, u is smooth on a data x space and a parameter θ space

3. Some applications of HGM

https://www.math.kobe-u.ac.jp/OpenXM/Math/hgm/ref-hgm.html

$\boxed{\text{openxm hgm}}$ search.

1. Finding the integration ideal $(I + \partial_{m+1} D_n + \cdots + \partial_n D_n) \cap D_m$: by hand (theoretical consideration) or by a new efficient algorithm.
2. Numerical algorithms to solve an ordinary differential equations (of huge size)[6].

[6] https://arxiv.org/abs/2111.10947

Fisher-Bigham distribution ($\supset$ Von-Mises distributuion)

$\frac{1}{Z(x,y,r)} \exp\left(\sum_{1\le i\le j\le d+1} x_{ij} t_i t_j + \sum_{i=1}^{d+1} y_i t_i\right) |dt|$ where Z is the normalizing constant

$$Z(x, y, r) = \int_{S^d(r)} \exp\left(\sum_{1\le i\le j\le d+1} x_{ij} t_i t_j + \sum_{i=1}^{d+1} y_i t_i\right) |dt|. \qquad (6)$$

- H.Nakayama et al, Holonomic Gradient Descent and its Application to Fisher-Bingham Integral (2011)[7].
- A.Kume, T.Sei, On the exact maximum likelihood inference of Fisher–Bingham distributions using an adjusted holonomic gradient method (2018)[8]
- S.Matsui, Finding initial values for MLE of the Fisher-Bingham distribution by a neural network (Kobe Univ. master thesis), 2024.

[7] https://doi.org/10.1016/j.aam.2011.03.001

[8] https://doi.org/10.1007/s11222-017-9765-3

Wishart distribution

Let X_i be a random column vector whose distribution is the m-dimensional multivariate normal (or Gaussian) distribution with mean vector 0 and covariance matrix Σ. $X = (X_1, \ldots, X_n)$.

Theorem 3 (Constantine (1963))

Let ℓ_1 be the maximal eigenvalue of $W = XX^T$. Then the probability that ℓ_1 is smaller than x is

$$P[\ell_1 < x] = C \exp\left(-\frac{x}{2}\mathrm{Tr}\,\Sigma^{-1}\right) x^{\frac{1}{2}nm} {}_1F_1\left(\frac{m+1}{2}; \frac{n+m+1}{2}; \frac{x}{2}\Sigma^{-1}\right) \tag{7}$$

Here,

$${}_1F_1(a; c; Y) = \frac{\Gamma_m(b)}{\Gamma_m(a)\Gamma_m(c-a)} \int_{0<X<I_m} \exp(\mathrm{Tr}\,XY)|X|^{a-(m+1)/2}|I_m - X|^{c-a-(m+1)/2} dX \tag{8}$$

is a holonomic function.

WishartHGM

```
#install.packages("hgm")
library("hgm")
hgm.pwishart(m=3,n=5,beta=c(1,2,3),q=3)
 [1] 3.0242949 0.5247871 ...  # it means P(ell<3.024)=0.524
plot(hgm.pwishart(m=3,n=5,beta=c(1,2,3),q=10,autoplot=1))
```

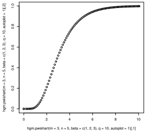


Figure: A graph of $P(\ell_1 < x)$

- Holonomic gradient method for the distribution function of the largest root of a Wishart matrix (2013)[9]
- H.Hashiguchi et al, Distribution of Ratio of two Wishart Matrices and Evaluation of Cumulative Probability by Holonomic Gradient Method (2018)[10],

[9] https://dx.doi.org/10.1016/j.jmva.2013.03.011

[10] https://doi.org/10.1016/j.jmva.2018.01.002

Neural tangent kernel

$f(\theta, x) \;:\; \mathbb{R}^{d_0} \overset{\text{am}}{\to} \mathbb{R}^{d_1} \overset{\sigma}{\to} \mathbb{R}^{d_1} \overset{\text{am}}{\to} \mathbb{R}^{d_2} \overset{\sigma}{\to} \mathbb{R}^{d_2} \to \cdots \to \mathbb{R}^{d_{L+1}}$,
"am"'s are affine maps with parameter θ. σ is an activation.

Theorem 4 (Jacot et al 2018[11])

When width d_i of neural network (NN) goes to ∞, the neural tangent kernel (NTK) $\left\langle \frac{\partial f(\theta,x)}{\partial\theta}, \frac{\partial f(\theta,x')}{\partial\theta} \right\rangle$ converges to $\Theta(x, x')$ in probability w.r.t. θ.

[11] `https://arxiv.org/abs/1806.07572`

$$f(x) \sim (\Theta(x, x_1), \Theta(x, x_2), \ldots, \Theta(x, x_N))(H^*)^{-1}(y_1, y_2, \ldots, y_N)^T. \tag{9}$$

$H^* = (\Theta(x_i, x_j))$ where x_i is input and y_i is output. Definition of Θ:

$$\Sigma^{(0)}(x, x') = x^T x', \tag{10}$$

$$\Lambda^{(h)}(x, x') = \begin{pmatrix} \Sigma^{(h-1)}(x, x) & \Sigma^{(h-1)}(x, x') \\ \Sigma^{(h-1)}(x', x) & \Sigma^{(h-1)}(x', x') \end{pmatrix} \tag{11}$$

$$\Sigma^{(h)}(x, x') = c_\sigma E_{(u,v)\sim N(0,\Lambda^{(h)})}[\sigma(u)\sigma(v)] \tag{12}$$

$$\dot{\Sigma}^{(h)}(x, x') = c_\sigma E_{(u,v)\sim N(0,\Lambda^{(h)})}[\dot{\sigma}(u)\dot{\sigma}(v)] \tag{13}$$

$$\Theta(x, x') = \Theta^{(L)}(x, x') = \sum_{h=1}^{L+1} \left(\Sigma^{(h-1)}(x, x') \prod_{h'=h}^{L+1} \dot{\Sigma}(x, x') \right) \tag{14}$$

Dual activation

$E_{(u,v)\sim N(0,\Lambda^{(h)})}[\sigma(u)\sigma(v)]$ (dual activation of σ)

$$\hat{E}[\sigma(u)\sigma(v)] = \int_{\mathbb{R}^2} \sigma(u)\sigma(v)\exp(x_{11}u^2 + 2x_{12}uv + x_{22}v^2)dudv$$

$$E_{(u,v)\sim N(0,\Lambda^{(h)})}[\sigma(u)\sigma(v)] = \hat{E}[\sigma(u)\sigma(v)]\frac{\sqrt{\det(x)}}{\pi}, \quad \Lambda^{(h)} = -\frac{1}{2}x^{-1}.$$

- ReLU (rectified linear unit)[12]: $\sigma(u) = uY(u)$.
- GeLU (Gaussian error linear unit): $\sigma(u) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{u}{\sqrt{2}}\right)\right)$

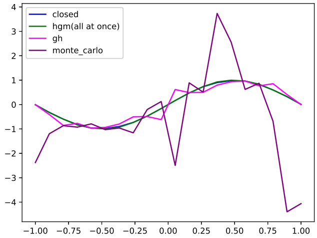
Evaluation of the dual holonomic activation can be perfomed by a HGM type algorithm.

A.Sakoda, N.Takayama, An Application of the Holonomic Gradient Method to the Neural Tangent Kernel, 2024[13].

[12] https://en.wikipedia.org/wiki/Activation_function

[13] http://arxiv.org/abs/2410.23626

Our algorithm ulitizes holonomic system for the expectation w.r.t. Gaussian distribution by T.Koyama and A.Takemura[14] and the restriction algorithm of T.Oaku[15].
Example: $\sigma(u) = Y(u)\sin(u)$. Interpolation by NTK Θ of $\sin(\pi x)$ with values at 15 points.



[14] https://doi.org/10.1007/s13160-015-0166-8, Calculation of Orthant Probabilities by the Holonomic Gradient Method (2015)
[15] https://doi.org/10.1006/aama.1997.0527 Algorithms for *b*-function, restrictions, and algebraic local cohomology groups (1997).

Summary

1. Holonomic functions or distributions are nice class of functions.
2. We can apply algebra and computer algebra to evaluate them (HGM).
3. MLE with respect to a holonomic unnormalized distribution can be described by a dynamical system.
4. HGM is applied to optimization, statistics, neural tangent kernel, physics, ...