Holonomic Decent Minimization Method for Restricted Maximum Likelihood Estimation

Rieko Sakurai[1] , and Toshio Sakata [2]

[1] *Graduate School of Medicine, Kurume University 67 Asahimachi, Kurume 830-0011, JAPAN*
[2] *Faculty of Design Human Science, Kyushu University, 4-9-1 Shiobaru Minami-ku, Fukuoka 815-8540, JAPAN*
email: a213gm009s@std.kurume-u.ac.jp

Abstract

Recently, the school of Takemura and Takayama have developed a quite interesting minimization method called *holonomic gradient descent method* (HGD). It works by a mixed use of Pfaffian differential equation satisfied by an objective holonomic function and an iterative optimization method. They successfully applied the method to several maximum likelihood estimation (MLE) problems, which have been intractable in the past. On the other hand, in statistical models, it is not rare that parameters are constrained and therefore the MLE with constraints has been surely one of fundamental topics in statistics. In this paper we develop HGD with constraints for MLE .

Key Words : Holonomic gradinet descent method, Newton-Raphson method with penalty function, von Mises-Fisher distribution

1 Introduction

Recently, the both schools of Takemura and Takayama have developed a quite interesting minimization method called holonomic gradient descent method(HGD). It utilizes Gröbner basis in the ring of differential operator with rational coefficients. Gröbner basis in the differential operators plays a central role in deriving some differential equations called a Pfaffian system for optimization. HGD works by a mixed use of Pfaffian system and an iterative optimization method. It has been successfully applied to several maximum likelihood estimation (MLE) problems, which have been intractable in the past. For example, HGD solve numerically the MLE problems for the von Mises-Fisher distribution and the Fisher-Bingham distribution on the sphere (see, Sei et al.(2013) and Nakayama et al.(2011)). Furthermore, the method has also been applied to the evaluation of the exact distribution function of the largest root of a Wishart matrix, and it is still rapidly expanding the area of applications(see, Hashiguchi et al.(2013)). On the other hand, in statistical models, it is not rare that parameters are constrained and therefore the MLE problem with constraints has been surely one of fundamental topics in statistics. In this paper, we develop HGD for MLE problems with constraints, which we call the constrained holonomic gradient descent(CHGD). The key of CHGD is to separate the process into (A) updating of new parameter values by Newton-Raphson method with penalty function and (B) solving a Pfaffian system.

2 Constrained Optimization Problem

We consider the following the constrained optimization problem.

$$\begin{aligned} (P) \quad & min \quad f(x) \\ & s.t. \quad g_i(x) \le 0,\ h_i(x) = 0 \end{aligned} \tag{1}$$

where $i = 1,...,m$, $j = 1,...,l$ and $f, g_i, h_j : R^n \to R$ are all assumed to be continuously differentiable function. $g_i(x)$ is an equality constraint function and $h_j(x)$ is an inequality constraint function. In this paper, the objective function $f(x)$ is assumed to be holonomic. We call the interior region defined by the constraint functions *the feasible region*.

2.1 Penalty Function Method

A penalty function method replaces a constrained optimization problem by a series of unconstrained problems. It is performed by adding a term to the objective function that consists of a penalty parameter ρ and a measure of violation of the constraints. In our simulation, we use *the exact penalty function method*. The definition of the exact penalty function is given by (see Yabe (2006)).

$$P(x;\rho) := f(x) + \rho\{\sum_{i=1}^{m} |g_i(x)| + \sum_{j=1}^{l} max(0, h_j(x))\},\ \rho > 0 \tag{2}$$

3 Holonomic descent method

Assume that we seek the minimum of a holonomic function $f(x)$ and the point x which gives the minimum $f(x)$. In HGD, we use the iterative method together with a Pfaffian system. In this paper, we use the the Newton-Raphson iterative minimization method in which the renewal rule of the search point is given by

$$x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k),$$

where $\nabla f(x_k) = (\frac{\partial f(x_k)}{\partial x_1}, \ldots, \frac{\partial f(x_k)}{\partial x_k})^T$ and $H(x_k)$ is the Hessian of $f(x)$ at $x = x_k$.

3.1 Mathematical background

HGD is based on the theory of the Gröbner basis. In the following, we refer to the relation of a numerical method and the Gröbner basis.

Let R be the differential ring written as

$$R = \boldsymbol{C}[x_1,...,x_n]\langle\partial_1,..,\partial_n\rangle$$

where $\boldsymbol{C}[x_1,...,x_n]$ are the rational coefficients of differential operators. Suppose that $I = \{\ell_i | i = 1,...,p\}$ is a left ideal of R, $k[\boldsymbol{x}]$ is a field and $D \in k[\boldsymbol{x}]\langle\partial_1,..,\partial_n\rangle \in I$. If an arbitrary function f satisfies $Df = 0$ for all D, then f is a solution of I. That is

$$\ell_i f = 0\ \forall i \tag{3}$$

When f satisfies Equation (3), f is called *holonomic function*.

Let $s = [s_1, \ldots, s_t]$, with $s_i = q_i(x)\partial^{\alpha_i}$ be a standard basis in the quotient vector space R/I which is a finite dimensional vector spaces. Let G be the Gröbner basis of I. The rank of arbitrary differential operations can be reduced by normalization by G. Assume that $\partial_i s_j \to_G \sum_k c^i_{jk} s_k$ holds. For a solution f of I put $F = (f, s_2 f, \ldots, s_t f)^T$. Then, it holds that

Proposition 3.1 *(see, e.g.,Nakayama et al.(2011))*

$$\frac{\partial F}{\partial x_i} = P_i F, \ \ i = 1, ..., n \tag{4}$$

where P_i is a $t \times t$ matrix with c^i_{jk} as a (j,k) element

Proof 3.1

$$\begin{aligned}
\frac{\partial s_j f(x)}{\partial x_i} &= (\partial_i \bullet s_j) f(x) \\
&= (\sum_k c^i_{jk} s_{jk}) f(x) \ \ (mod \ \ I) \\
&= [P_i F]_j, \ \ i = 1, ..., n, \ \ j = 1 ..., t
\end{aligned} \tag{5}$$

This proves the assertion.

The above differential equations are called *Pfaffian differential equations* or *Pfaffian system* of I. So we can calculate the gradient of F by using Pfaffian differential equations. Then, $\nabla f(x_k)$ and $H^{-1}(x_k)$ are also given by Pfaffian differential equations. (see Hibi et al.(2012))

Lemma 3.1 *Let $\sum_j^t a_{ij} s_j$ be the normal form of $\partial_i = \partial/\partial x_i$ by G and $\sum_k^t u_{ijk} s_k$ be the normal form of $\partial_i \partial_j$ by G. Then we have,*

$$\partial_i f(x_k) \to_G (\sum_j^t a_{ij} s_j) f(x_k) = \sum_j^t a_{ij} F_j(x_k) = ((P_1 F(x_k))_1, ..., (P_n F(x_k))_1) \tag{6}$$

$$\partial_i \partial_j f(x_k) \to_G (\sum_m^t u_{ijk} s_m) f(x_k) = \sum_m^t u_{ijm} F_m(x_k) = ((\frac{\partial P_i}{\partial x_j} + P_i P_j) F(x_k))_1 \tag{7}$$

where $(v)_1$ denotes the first entry of a vector v.

3.2 Algorithm

For HGD, we first give an ideal $I = \{\ell_i | i = 1, ..., p, \ \ell_i f = 0 \ \forall i\}$ for holonomic function $f(x)$ and calculate the Gröbner basis G of I and then the standard basis S are given by G. The coefficient matrix P_i for Pfaffian system is led by this standard basis, and $H^{-1}(x_k)$ and $\nabla f(x_k)$ are calculated from S by starting from a initial point x_0 through the Pfaffian equations. After these, we can compute automatically the optimum solution by a mixed use of then Newton-Raphson method. The algorithm is given by below.

Algorithm 3.1

step 1 Set $k = 0$ and take an initial point x_0 and evaluate $F(x_0) = (f(x_0), s_1 f(x_0), \ldots, s_t f(x_0))^T$.

step 2 Evaluate $\nabla f(x_k)$ and $-H^{-1}(x_k)$ from F and calculate the Newton direction, $d_k = -H^{-1}(x_k)\nabla f(x_k)$

step 3 Update a search point by $x_{k+1} = x_k + d_k$.

step 4 Evaluate $F(x_{k+1})$ by solving Pfaffian equations numerically.

step 5 Set $k = k+1$ and calculate $F(x_{k+1})$ and goes to step.2 and repeat until convergence.

Remark 3.1 *The key step of the above algorithm is step 4. We can not evaluate $F(x_{k+1})$ by inputting x_{k+1} in the function $f(x)$ since the HGD treats the case that $f(x)$ is difficult to calculate numerically. Instead, we only need calculate $f(x_0)$ and $F(x_0)$ numerically for a given initial value x_0.*

4 Constrained holonomic gradient descent method

Now, we propose the method in which we add constraint conditions to HGD and call it the constrained holonomic gradient descent method(CHGD).

4.1 How to add the constraints

For treating constraints we use the penalty function and add it to objective function and make a new objective function and can treat it as the unconstrained optimization problem. We use HGD for evaluation of gradients and Hessian and use the exact penalty function method for constraints. The value of updating a search point can be obtained as the product of directional vector and step size. The step size α is chosen so that the following Armijo condition is satisfied. In fact we chose α such that

$$P(x_k + \alpha x_k; \rho) \leq P(x_k; \rho) + \xi \alpha \{P_l(x_k, \nabla x_k; \rho) - P(x_k; \rho)\}, \tag{8}$$

where $0 < \xi < 1$ and $P_l(x_k, \nabla x_k; \rho)$ is the approximation of $P(x_k, \rho)$ given by.

$$\begin{aligned} P_l(x_k, \nabla x_k; \rho) &= f(x_k) + \nabla f(x_k)^T \nabla x_k \\ &+ \rho\{\sum_{i=1}^{m} |g_i(x_k) + \nabla g_i(x_k)^T \nabla x_k| \\ &+ \sum_{j=i}^{n} max(0, h_j(x_k) + \nabla h_j(x_k)^T \nabla x_k)\} \end{aligned} \tag{9}$$

The initial value of α is set 1 and then α is made smaller iteratively until α satisfies Equation (8), or $\alpha = 0$.

In our algorithm, holonomic gradient descent plays a role to calculate the gradient vectors and then the penalty function plays a role to control the step size iteratively.

5 Computational results

We apply CHGD for MLE for von Mises distribution(vM). The process of applying for HGD is shown in Nakayama et al.(2011). The density function of vM is given by $f(\kappa,\mu) = e^{\kappa\cos(\mu-x)}/\int_0^{2\pi} e^{\kappa\cos(\mu-t)}dt$. The parameters of vM, κ and μ, show concentration and mean of angle data $\boldsymbol{x}$ respectively. We set the parameters for MLE $\theta_1 = \kappa\cos\mu$ and $\theta_2 = \kappa\sin\mu$. Now we solve the constrained optimization problem given by.

$$\begin{aligned} (P) \quad min \quad & L(\theta_1,\theta_2) = e^{-\bar{c}\theta_1-\bar{s}\theta_2}\int_0^{2\pi} e^{\theta_1\cos t+\theta_2\sin t}dt \\ s.t. \quad & \theta_1 \le \theta_2 \end{aligned} \tag{10}$$

Let $\boldsymbol{x}$ be sample data. Let n be sample size. Then, $\bar{c} = \frac{1}{n}\sum_i^n \cos x_i$ and $\bar{s} = \frac{1}{n}\sum_i^n \sin x_i$.

5.1 Simulation

In our simulation, we set the vM's parameter $(\kappa,\mu) = (5,\pi/4)$ of which the true value $(\theta_1,\theta_2) = (3.54,3.54)$ and the initial value $(\theta_1,\theta_2) = (-2.0,0.1)$. We tried the 2 patterns of constraints. Both of the case worked under the same condition except constraints. In Figure 1, the constraint is $\theta_1 \le \theta_2$. In Figure 2, the constraint is $\theta_1^2+\theta_2^2 \le 9$. Figures 1,2 are the drawing of the trace of the search point.

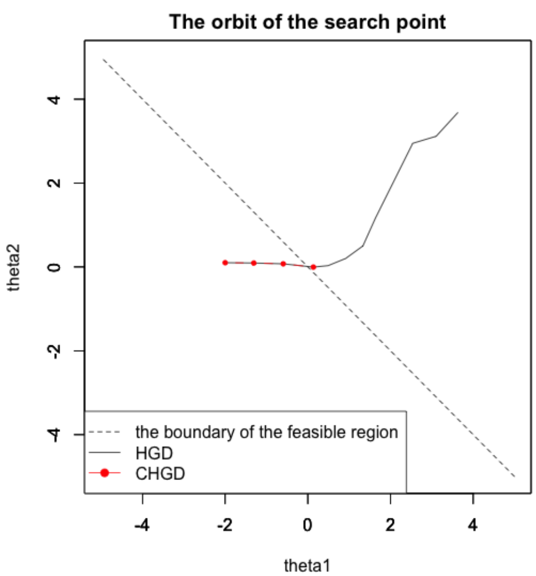

Figure 1: the case of $\theta_1 \le \theta_2$

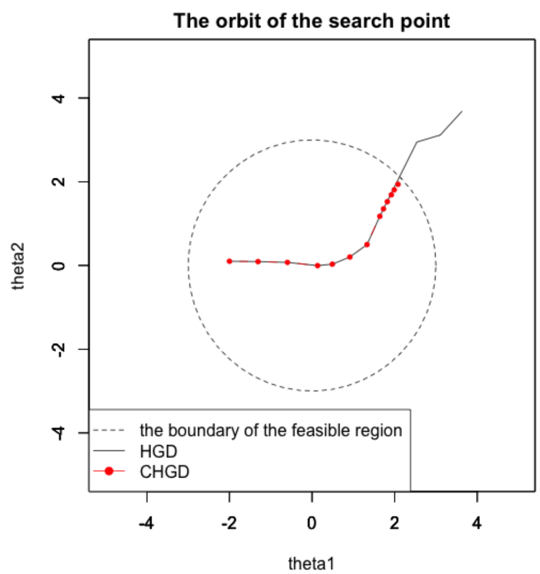

Figure 2: the case of $\theta_1^2+\theta_2^2 \le 9$

The result of simulation, the convergence point of HGD is $(\theta_1,\theta_2) = (3.63,3.67)$. In Figure 1, the convergence point of CHGD is $(\theta_1,\theta_2) = (0.13,-0.004)$. In Figure 2, the convergence point of CHGD is $(\theta_1,\theta_2) = (2.08,1.94)$. In the CHGD, the search direction is almost same as the HGD, because the direction is decided by the HGD's algorithm. While, the constraints play the role to judge the search point is within the feasible region or not and decide the step size.

5.2 Runtimes

CHGD is the effective method for optimization with constraints. However, whenever CHGD increases the cost of runtimes than HGD regardless of whether the solution is in the feasible region or not. The following table shows the runtimes when the optimization solution is within the feasible region.

Table 1: Comparing the runtimes

	CPU TIME (sec)	PATAMETERS (θ_1, θ_2)	
HGD	0.03698	2.120627	2.120333
CHGD	0.09834	2.120803	2.120629
NEWTON	0.12598	2.124429	2.124855

We programmed by R and executed on Windows 7 64bit with RStudio Version 0.97.336

In Table 1, all numbers are the means of 500 times trials. The optimization problem is Equation (10). Sample data is drawn from the vM with $(\theta_1, \theta_2) = (2.12, 2.12)$. The third column of Table 1 is the result with only Newton-Raphson method which optimize $f(x)$ directly, not use Pfaffian system. Thus, we see that HGD and CHGD is faster than Newton-Raphson method.

We see that the runtimes of CHGD is longer than HGD in general, where the both of solutions are almost the same value when the solution is inside the feasible region. Sometimes the process finishes early by constraints, when the solution is outside the feasible region. Although, we need consider the cost of calculation of CHGD.

References

[1] Hashiguchi, H., Numata, Y., Takayama, N., Takemura, A. (2013). *"The holonomic gradient method for the distribution function of the largest root of a Wishart matrix"*. Journal of Multiva, riate Analysis 117 (2031) 296-312

[2] Sei, T., Shibata, H., Takemura, A., Ohara, K., Takayama, N. (2013). *"Properties and applications of Fisher distribution on the rotation group"*. Journal of Multivariate Analysis.

[3] Hibi, T., Hamada, T., Noro, M., Aoki, S., Takemura, A., Osugi, H., Takayama, N., Nakayama, H., Nishiyama, K. (2012). *"Gröebner dojo(japanese)"*. Kyoritsu publisher.

[4] Nakayama, H., Nishiyama, K., Noro, M., Ohara, K., Sei, T., Takayama, N., Takemura, A. (2011). *"Holonomic gradient descent and its application to the Fisher–Bingham integral"*. Advances in Applied Mathematics, 47(3), 639-658.

[5] Yabe, H. (2006). *"Introduction and Application of Optimization Problem(japanese)"*. Surikougakusha publisher.

[6] Cox, D. A., Little, J., O'Shea, D. (2007). *"Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra (Vol. 10)"*. Springer Verlag.