GRADIENT FLOWS ON A SPACE OF POWER-FUNCTIONS FOR MINIMIZING A CONVEX FUNCTIONAL

YUKIO HAYASHI*
Japan Advanced Institute of Science and Technology, Hokuriku
Tatsunokuchi 923-1292, Ishikawa, Japan
E-mail:`yhayashi@jaist.ac.jp`

Keywords: optimization algorithm, self-concordant barrier function, information geometry, gradient system.

2000 Mathematics Subject Classification: *Primary: 90C25; Secondary: 53C22, 90C51.*

1 Introduction

We study a class of nonlinear dynamical systems to develop efficient algorithms. As an efficient algorithm, interior point method based on Newton's method is well known for solving convex programming (CP) problems which include linear, quadratic, semidefinite, and l_p-programming problems [9] [13]. On the other hand, the geodesic of information geometry [1] [2] is represented by a continuous Newton's method for minimizing a convex functional called divergence. Thus, we discuss a relation between information geometry and CP in a related family of continuous Newton's method. In particular, as the optimization of parameter values, we consider the α-projection problem from a given data onto an information geometric submanifold spanned with power-functions such as a weighted l_p-norm [11]. Originally, information geometry was constructed to introduce a natural structure for a family of probability distributions over continuous variables in statistical theory [1]. It has been successfully applied [2] with a convex function and a Legendre transformation in many areas: system theory, information theory, neural networks, quantum physics, mathematical programming and integrable systems.

First, as a deeply relation, we present there exists a structural similarity between the α-projection and semidefinite programming (SDP) problems [14]. Both problems are solvable by following the path on geodesics. The geometric structure is based on the autoparallelisms [1] [2] or linear property in the function space over finite discrete variables or the space of positive definite matrices, respectively. The property is practically applied to derive approximation methods for iteratively calculating the geodesic. The maximum step-length is determined by geometric quantities with respect to the Riemannian metric and the dual-connection. We show the proposed method is effective in a simulation.

Next, we reconsider the α-projection problem as a l_p-programming and the related ones, and reformulate it into a form of CP. From the reformulated problems, we derive self-concordant barrier functions [9] [13] according to the real values of α. It means the existence of a polynomial time interior-point algorithm for our problem.

Furthermore, we present the coincidence with the gradient directions on the geodesic for the divergence and on the affine-scaling (AS) trajectory for a modified barrier function. In such class of dynamical systems related to Newton's method, these results connect part of nonlinear and algorithmic analyses with the discreteness of variables.

2 Geodesic on the α-affine manifold

Let us consider a m-dimensional manifold $\tilde{M} \stackrel{\text{def}}{=} \{\tilde{f}(x) |\ 0 < \tilde{f}(x) < \infty, x \in \mathcal{X}\}$ consists of finitely bounded functions over $\mathcal{X} \stackrel{\text{def}}{=} \{1, 2, \dots, m\}$. By the discreteness, integral with respect to $x \in \mathcal{X}$ can be more easily treated as summation or vector-matrix operation. Introducing the primal parameter $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_n)^T$, we define a n-dimensional submanifold $\tilde{S}_\alpha = \{m_\alpha(x;\ \tilde{\theta})\}$ called α-affine manifold [1] [2],

$$m_\alpha(x;\ \tilde{\theta}) \stackrel{\text{def}}{=} \left[\frac{1-\alpha}{2}\sum_{i=1}^{n} F_i(x)\tilde{\theta}^i\right]^{\frac{2}{1-\alpha}}, \tag{1}$$

where we assume $m_\alpha(x;\ \tilde{\theta}) > 0$, $\sum_{x\in\mathcal{X}} m_\alpha(x;\ \tilde{\theta}) < \infty$, $n < m$ and $\alpha \in \mathbf{R}$ $(\alpha \neq \pm 1)$. Although $[F_i(x)]$ is a $m\times n$ matrix because of the discreteness of variable x, we use the notation to emphasis the meaning of function. Each basis function $F_i(x)$ is linearly independent to let the Fisher information metric exist as

$$g_{ij}(\tilde{\theta}) \stackrel{\text{def}}{=} \partial_i\partial_j\psi(\tilde{\theta}) = \sum_{x\in\mathcal{X}} \partial_i \ln m_\alpha \cdot \partial_j \ln m_\alpha \cdot m_\alpha, \tag{2}$$

where we denote $\partial_i \stackrel{\text{def}}{=} \partial/\partial\tilde{\theta}^i$, and $\psi(\tilde{\theta})$ is a convex function

$$\psi(\tilde{\theta}) = \frac{2}{1+\alpha}\sum_{x\in\mathcal{X}} m_\alpha(x;\ \tilde{\theta}).$$

Now, we will discuss a parameter estimation problem called α-projection as shown in Fig. 1. The problem is defined by minimizing a convex functional called α-divergence $D_\alpha(m_\alpha||\tilde{q})$,

$$D_\alpha(m_\alpha||\tilde{q}) \stackrel{\text{def}}{=} \frac{4}{1-\alpha^2}\left[\frac{1-\alpha}{2}m_\alpha + \frac{1+\alpha}{2}\tilde{q} - m_\alpha^{\frac{1-\alpha}{2}}\tilde{q}^{\frac{1+\alpha}{2}}\right]. \tag{3}$$

This is a quasi-distance between the function $m_\alpha(x;\ \tilde{\theta})$ on $\tilde{S}_\alpha$ and a given data $\tilde{q}(x)$ in $\tilde{M}$. In an applicational point of view, this problem is an estimation on the power-functions corresponded to fuzzy averaging operators [3] [12] between AND-OR logic according to the values of α.

As the information geometric structure [1] [2] of $\tilde{S}_\alpha$, the minimization of D_α is solvable by a straight line in the dual coordinate system $[\tilde{\eta}_i]$, which is one-to-one corresponded to $[\tilde{\theta}^i]$ with the Legendre transformation

$$\varphi(\tilde{\eta}) = \sum_{i=1}^{n}\tilde{\theta}^i\tilde{\eta}_i - \psi(\tilde{\theta}), \tag{4}$$

where $\varphi(\tilde{\eta})$ is another convex function. Because we have the optimal condition to minimize D_α,

$$\partial_j D_\alpha = \frac{2}{1+\alpha}\left\{\sum_x F_j m_\alpha^{\frac{1+\alpha}{2}} - \sum_x F_j \tilde{q}^{\frac{1+\alpha}{2}}\right\} = \tilde{\eta}_j - \tilde{\eta}_j(T) = 0, \tag{5}$$

using the definitions $\tilde{\eta}_j(T) \stackrel{\text{def}}{=} \frac{2}{1+\alpha}\sum_x F_j\tilde{q}^{\frac{1+\alpha}{2}}$ and

$$\tilde{\eta}_j \stackrel{\text{def}}{=} \frac{2}{1+\alpha}\sum_x F_j \tilde{m}_\alpha^{\frac{1+\alpha}{2}},$$

from (3). We should remark $\tilde{\eta}_j = \partial_j\psi(\tilde{\theta})$ and $g_{ij} = \partial_i\partial_j\psi(\tilde{\theta}) = \partial_i\tilde{\eta}_j$ by the differentiation of (4).

Thus, the estimated point is given by the following equations at $t' \to \infty$ and $t = 1$.

Exponential time scale ($0 \le t' < \infty$):

$$\frac{d\tilde{\eta}_j}{dt'} = -(\tilde{\eta}_j - \tilde{\eta}_j(T)).$$

Linear time scale ($0 \le t \le 1$):

$$\frac{d\tilde{\eta}_j}{dt} = -(\tilde{\eta}_j(I) - \tilde{\eta}_j(T)) \stackrel{\text{def}}{=} \Delta\tilde{\eta}_j :\ const. \tag{6}$$

Since they represent a straight line in the coordinate system $[\tilde{\eta}_i]$, it is a (- α)-geodesic [1] [2] between any initial and terminal points† . The solutions

$$\tilde{\eta}_j = \tilde{\eta}_j(T) + (\tilde{\eta}_j(I) - \tilde{\eta}_j(T))e^{-t'},$$

$$\tilde{\eta}_j = (1-t)\tilde{\eta}_j(I) + t\tilde{\eta}_j(T), \tag{7}$$

are equivalent within the time-scale transformation $1 - t = e^{-t'}$. Both trajectories converge to the terminal point $\tilde{\eta}(T)$ on the same straight line with different speeds in the n-dimensional $[\tilde{\eta}_i]$, while the linearity is more clear in (7).

On the other hand, the geodesic is also represented by the gradient system [4]:

$$\frac{d\tilde{\theta}^i}{dt'} = -\sum_{j=1}^{n} g^{ij}(\tilde{\theta})\partial_j D_\alpha(m_\alpha || \tilde{q}), \tag{8}$$

where $[g^{ij}]$ denotes the inverse matrix of $[g_{ij}]$.

Since the Hessian $\partial_i \partial_j D_\alpha = g_{ij}$ is derived from (1), (3) and (4), the gradient system (8) is nothing but continuous Newton's method for minimizing the α-divergence D_α. However, the iterative calculation of (8) such as Runge-Kutta approximation requires much computations in the inverse matrix at each iteration. In general, the inverse transformation from $\tilde{\eta}$ to $\tilde{\theta}$ is implicit, we must solve the gradient system (8).

If $\tilde{S}_\alpha$ is $\pm\alpha$-autoparallel in $\tilde{M}$, then the m-dimensional vector $m_\alpha^{\frac{1+\alpha}{2}}$ is linear, and the value of $\tilde{\theta}^i$ is directly solvable [7] without any iterations of the gradient system (8). Note that the α-autoparallelism is trivial [1] [2], because an α-geodesic on $\tilde{S}_\alpha$ is a straight line in $[\tilde{\theta}^i]$, and also a straight line of $m_\alpha^{\frac{1-\alpha}{2}}$ in $\tilde{M}$ from (1). In this $\pm\alpha$-autoparallel case, at any points on the geodesic, we can linearly interpolate $m_\alpha^{\frac{1+\alpha}{2}}$ between the initial $m_\alpha(x;\ \tilde{\theta}(I))$ and the terminal $m_\alpha(x;\ \tilde{\theta}(T))$ corresponded to $\tilde{\theta}(I)$ and $\tilde{\theta}(T)$. Therefore, by substituting the interpolated value of $m_\alpha^{\frac{1+\alpha}{2}}$ into the definition (1), we have a system of linear equations in the vector-matrix form

$$F\tilde{\theta} = \frac{2}{1-\alpha} m_\alpha^{\frac{1-\alpha}{2}}. \tag{9}$$

Once the QR factorization of F is numerically obtained (e.g. by using modified Gram-Schmidt algorithm [5]), as a computational merit, it can be commonly applied to the other initial, terminal, and the interpolated values of $m_\alpha^{\frac{1+\alpha}{2}}$.

3 Parameter estimation onto the submanifold

When an observed data $\tilde{q}$ in $\tilde{M}$ is outside from the submanifold $\tilde{S}_\alpha$, a $(-\alpha)$-geodesic is applied to the parameter estimation problem. Based on the minimization of the α-divergence D_α, the orthogonally projected point onto $\tilde{S}_\alpha$ from $\tilde{q}$ is given by the convergent point of the gradient system (8) as the geodesic from any initial point. The estimated function m_α satisfies the system of linear equations in the vector-matrix form

$$F^T m_\alpha^{\frac{1+\alpha}{2}} = \frac{1+\alpha}{2}\tilde{\eta}(T), \tag{10}$$

by the optimal condition (5).

Thus, as shown in [7], the estimation problem is formally reduced to double systems of linear equations (9) and (10). However, in (10), the linear mapping from the n-dimensional $\tilde{\eta}(T)$ to the higher m-dimensional variables $m_\alpha^{\frac{1+\alpha}{2}}$ has many solutions on the α-projection curve from $\tilde{q}(x)$ onto $\tilde{S}_\alpha$ in $\tilde{M}$. In other words, $\tilde{\eta}$ and $m_\alpha^{\frac{1+\alpha}{2}}$ are not one-to-one corresponding, there exists a null space $\{f |\ F^T f = 0\}$.

† In this notation $\tilde{\eta}(I)$ or $\tilde{\eta}(T)$, the parenthesis denotes the dependence on the initial or terminal point.

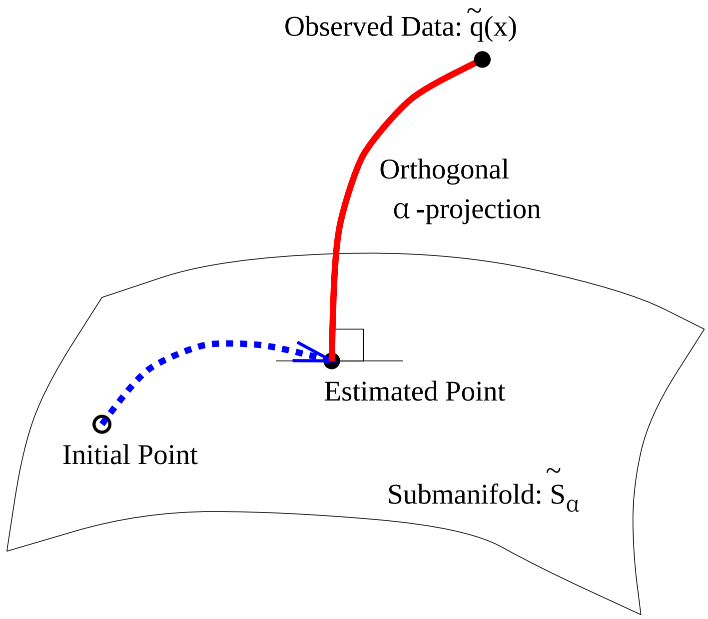

Figure 1: Parameter estimation from a given data $\tilde{q}(x)$ onto the α-affine manifold $\tilde{S}_\alpha = \{m_\alpha(x;\ \tilde{\theta})\}$.

3.1 Autoparallel case

If $\tilde{S}_\alpha$ is $\pm\alpha$-autoparallel, $m_\alpha^{\frac{1+\alpha}{2}}$ is linear on a $(-\alpha)$-geodesic (and also $m_\alpha^{\frac{1-\alpha}{2}}$ is linear on an α-geodesic). In this case, to numerically solve (10), we may apply a proper line search algorithm [15] in the m-dimensional space of $m_\alpha^{\frac{1+\alpha}{2}}$ for minimizing D_α. The search direction is determined by several steps of the gradient system (8). Then, we substitute the solution m_α for (5) into (9), and similarly apply the QR factorization of F to it [7]. This method has been proposed in the case of em-autoparallel exponential or mixture family [6].

The $\pm\alpha$-autoparallelism is based on the linearity of power-function $m_\alpha^{\frac{1+\alpha}{2}}$ in $\tilde{M}$, while the case of SDP [14] is based on the linearity of inverse matrix $P^{-1}(z)$ in the space of the positive definite matrices $PD(m)$. In spite of the quite different problem formulations, there exists a same structure as shown in Table 1.

3.2 General case with $KerF^T$

In general, since $\tilde{\eta}$ and $m_\alpha^{\frac{1+\alpha}{2}}$ are not one-to-one corresponding, we must consider $KerF^T$ for the mapping from $m_\alpha^{\frac{1+\alpha}{2}}$ to $\tilde{\eta}$ in (10). Although the n-dimensional $\tilde{\eta}$ is linear on the geodesic (7), the extended m-dimensional $m_\alpha^{\frac{1+\alpha}{2}}$ has higher terms of time variable t. It is represented by the following vector-matrix form

$$\frac{1+\alpha}{2}\tilde{\eta} = F^T m_\alpha^{\frac{1+\alpha}{2}} \stackrel{\text{def}}{=} F^T(f_0 + f_1 t) + \sum_{k\geq 2} F^T f_k t^k, \tag{11}$$

$$f_0,\ f_1 \notin KerF^T, \quad \sum_{k\geq 2} f_k t^k \in KerF^T,$$

where $f_0, f_1, f_2, \ldots$ are m-dimensional vectors, and $rank\ F^T + dim(Ker\ F^T) = n + (m-n) = m$.

If the second term in the right-hand side of (11) is small enough, the curved line of $m_\alpha^{\frac{1+\alpha}{2}}$ is approximated by the linear component $f_0 + f_1 t$. In the next two sections, as similar to a general case of nonlinear $P^{-1}(z)$ in SDP [14], we will propose two kinds of approximation algorithms, and confirm the effectiveness by a simulation.

	α-projection: min D_α	SDP: min $c^T z$
whole space (m-dimension)	$\tilde{M} = \{\tilde{f}(x)\}$ function space over $\mathcal{X}$	$PD(m)$ space of positive definite matrices
constraint submanifold (n-dimension, $n < m$)	$\tilde{S}_\alpha = \{m_\alpha(x;\, \tilde{\theta})\}$ $m_\alpha(x;\, \tilde{\theta}) = \left[\frac{1-\alpha}{2}\sum_{i=1}^n F_i(x)\tilde{\theta}^i\right]^{\frac{2}{1-\alpha}}$ $\{F_i(x)\}$: basis functions	$\mathcal{L} = PD(m) \cap \{P(z)\}$ $P(z) = E_0 + \sum_{i=1}^n E_i z^i \geq O$ $\{E_i\}$: space of symmetric matrices
parameters	$\tilde{\theta}$ and $\tilde{\eta}$ in $\tilde{S}_\alpha$ $m_\alpha^{\frac{1\mp\alpha}{2}}$ in $\tilde{M}$	z and y in $\mathcal{L}$ $y_i = -tr(P(z)^{-1}E_i)$ ξ and ζ in $PD(m)$ $P^{-1} = \sum_{i=1}^m \xi^i E_i = \left(\sum_{i=1}^m \zeta_i \acute{E}^i\right)^{-1}$
object function	α-divergence $D_\alpha(m_\alpha \vert\vert \tilde{q})$	self-concordant barrier $\Phi_t(z) = t \times c^T z + \phi(z)$
metric g_{ij}	$\partial_i\partial_j D_\alpha = \partial_i\partial_j\psi$ $\psi(\tilde{\theta}) = \frac{2}{1+\alpha}\sum_x m_\alpha(x;\, \tilde{\theta})$	$\partial_i\partial_j\phi$ $\phi(z) = -\log\det P(z)$
connections	$\pm\alpha$-geodesics	∇-, ∇^*-geodesics
linearity	linearities of $m_\alpha^{\frac{1\mp\alpha}{2}}$ $\pm\alpha$-autoparallel	linearities of $P^{\pm 1}$ ∇-∇^* autoparallel
search direction	$(-\alpha)$-geodesic Newton's method	∇^*-geodesic AS-trajectory

Table 1: Correspondences between the α-projection and SDP problems in the structural similarity.

4 Approximation methods for the geodesic

4.1 Piece-wise linear approximation

Let us consider piece-wise linear lines of $m_\alpha^{\frac{1+\alpha}{2}}$ in the m-dimensional function space $\tilde{M}$. At the current point t, we have the Taylor expansion

$$m_\alpha(x,\tilde{\theta}(t+\delta t))^{\frac{1+\alpha}{2}} = m_\alpha(x,\tilde{\theta}(t))^{\frac{1+\alpha}{2}} + \left.\frac{dm_\alpha^{\frac{1+\alpha}{2}}}{dt}\right|_t \delta t + \frac{1}{2}\left.\frac{d^2 m_\alpha^{\frac{1+\alpha}{2}}}{dt^2}\right|_t \delta t^2 + O(\delta t^3), \tag{12}$$

The first and second derivatives corresponded to f_1 and f_2 in (11) are

$$\frac{dm_\alpha^{\frac{1+\alpha}{2}}}{dt} = \sum_i \partial_i m^{\frac{1+\alpha}{2}} \frac{d\tilde{\theta}^i}{dt} = \frac{1+\alpha}{2} m^\alpha \sum_i F_i \frac{d\tilde{\theta}^i}{dt},$$

$$\frac{d^2 m_\alpha^{\frac{1+\alpha}{2}}}{dt^2} = \frac{1+\alpha}{2} m^\alpha \sum_{i,j}\left\{\frac{F_iF_j}{m^{\frac{1+\alpha}{2}}} - \sum_k F_k \Gamma_{ij}^k\right\}\frac{d\tilde{\theta}^i}{dt}\frac{d\tilde{\theta}^j}{dt},$$

where the $(-\alpha)$-connection coefficient [1] [2] is

$$\Gamma^*_{ijh}(\tilde{\theta}) \stackrel{\mathrm{def}}{=} \alpha \times \sum_x F_i(x)F_j(x)F_h(x)\left[m_\alpha(x;\,\tilde{\theta})\right]^{\frac{-1+3\alpha}{2}},$$

$$\Gamma^k_{ij}(\tilde{\theta}) = \sum_h g^{hk}(\tilde{\theta})\Gamma^*_{ijh}(\tilde{\theta}). \tag{13}$$

Remember that, from (6) and $g^{ij} = \frac{\partial\tilde{\theta}^i}{\partial\tilde{\eta}_j}$, the $(-\alpha)$-geodesic in the linear time-scale is derived as

$$\frac{d\tilde{\theta}^i}{dt} = \sum_j g^{ij}\frac{d\tilde{\eta}_j}{dt} = \sum_j g^{ij}\Delta\tilde{\eta}_j. \tag{14}$$

When the amount of the quadratic term in (12) is small, the value of $m_\alpha^{\frac{1+\alpha}{2}}$ is linearly approximated in the m-dimensional space. In the accuracy $|\frac{1}{2}\frac{d^2f}{dt^2}|\delta t^2 < \varepsilon^2$, the maximum step-length is obtained as

$$\delta t < \min_x \frac{2\varepsilon}{\sqrt{\left|(1+\alpha)m_\alpha(x;\,\tilde{\theta})^\alpha \sum_{i,j}\left\{\frac{F_i(x)F_j(x)}{m_\alpha(x;\,\tilde{\theta})^{\frac{1+\alpha}{2}}} - \sum_k F_k(x)\Gamma^h_{ij}\right\}\frac{d\tilde{\theta}^i}{dt}\frac{d\tilde{\theta}^j}{dt}\right|}}$$

Thus, we iteratively calculate the step-length δt by using each updated value of $m_\alpha^{\frac{1+\alpha}{2}}$ on the piece-wise linear line.

Step 0: Set an initial value $\tilde{\theta} = \tilde{\theta}(I)$ and an observed data $\tilde{q}(x)$.
Calculate the values of m_α, $\tilde{\eta}$, $\tilde{\eta}(T)$, $\Delta\tilde{\eta}$, g_{ij}, Γ^k_{ij}, and $\frac{d\tilde{\theta}^i}{dt}$.

Step 1: Calculate the maximum step-length δt.

Step 2: Approximate $m_\alpha^{\frac{1+\alpha}{2}}$ by the linear component $f_0 + f_1\delta t$.

Step 3: Update the value of g_{ij}, Γ^k_{ij}, $\frac{d\tilde{\theta}^i}{dt}$, and $\tilde{\eta}_j$ for the approximated m_α.

Step 4: If the current value of $\tilde{\eta}$ is sufficiently near to $\tilde{\eta}(T)$, then stop.
Otherwise, return to Step 1.

For the case of exponential or mixture family ($\alpha = \pm 1$), the maximum step-length δt is also derived in Appendix.

4.2 Predictor-corrector algorithm

Although it is expected that the above predictor method is efficient rather than the standard calculation method such as Runge-Kutta approximation for (8) or (14), the piece-wise linear line may slightly leave from the submanifold $\tilde{S}_\alpha$. The iterative process increases the accumulated error. In other words, the step-length needs to be set as considerably small.

For this problem, we consider the corrector process in which the value of m_α returns onto the submanifold. After Step 2, the approximated value of $\tilde{\theta}^*$ is obtained through the QR-factorization of F for (9) by the minimization of the 2-norm

$$\left\| F\tilde{\theta} - \frac{2}{1-\alpha}m_\alpha^{\frac{1-\alpha}{2}} \right\|^2.$$

The value of $\tilde{\theta}^*$ is applied into (1) as the corrected value $m_\alpha(x;\,\tilde{\theta}^*)$ on $\tilde{S}_\alpha$, then the update process in Step 3 is performed. We should remark that the obtained trajectory by the predictor-corrector algorithm is connected with the $(-\alpha)$-geodesics from different initial points, as shown in Fig. 2.

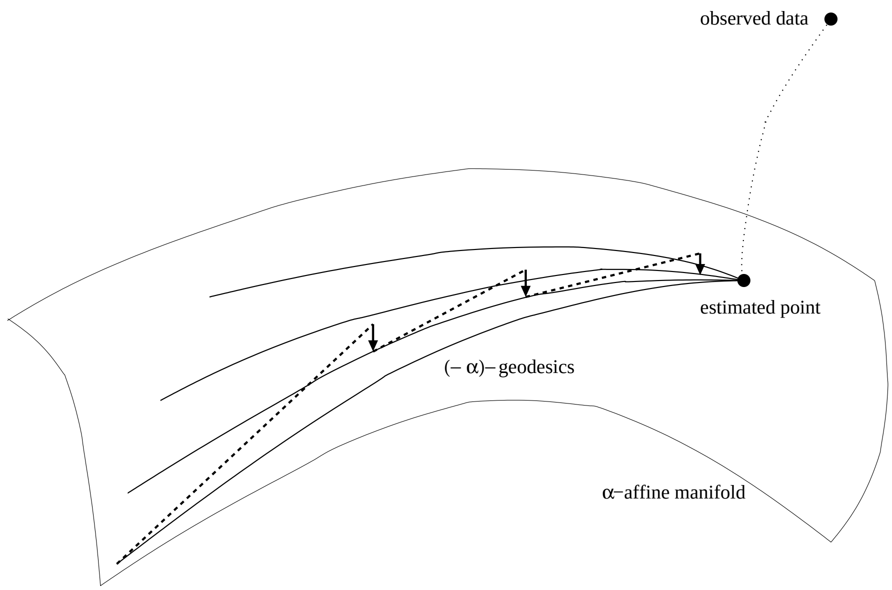

Figure 2: The piece-wise linear lines (dashed lines) and the corrector process (bold arrows) of the projection onto the submanifold $\tilde{S}_\alpha$.

5 Simulation result

By a simulation, we investigate the iteration times and the estimation error for the proposed methods. We set as, $n = 9$, $m = 125$, $\varepsilon^2 = 10^{-4}$, where the basis functions $F_i(x)$ and the observed data $\tilde{q}(x)$ consist of step-values from 1 to 5. In the piece-wise linear and predictor-corrector methods, the value of $\tilde{\theta}$ is iteratively updated from an initial random value $\tilde{\theta}(I)$, until the sum of step-lengths $\sum \delta t = 1$ corresponded as the terminal time $t = 1$ for (7) or (14).

First, we show the iteration times for these methods in Fig. 3. The reciprocal of the iteration time is the average step-length δt at each value of α. There are much more iterations for the piece-wise linear method, especially, the corrector method is effective in $\alpha > 1$. However, in both methods, many iterations are required as large α or $\alpha \approx 1$. We should note the case of $\alpha = 1$ known as exponential family in statistics is the best.

Second, we show the mean square error (MSE) in Fig. 4, where we omit the case of $\alpha = 1$ because only it has different normalized scale. There are numerical overflows with very large values of m_α and $\tilde{\theta}$ in the cases of $\alpha \approx 1$ for all of the three data. The phenomena is slightly similar to the non-robustness for a few data as outliers ($p \to \infty$: corresponding to $\frac{1-\alpha}{2} \to \infty$) on the L_p-norm $\|e\| \stackrel{\text{def}}{=} [\sum_i |e_i|^p]^{\frac{1}{p}}$ based on the p-th power of error elements $\{e_i\}$ in the linear inverse problem [11]. Here, both proposed methods are in the same error level at each value of α. The estimated values are continuously changed according to the value of α except in the case of $\alpha = 1$. These results suggest that the overflow is caused by the intrinsic estimation property in the families of the power-functions rather than by the numerical error. The reason will be discussed later. As shown in Table 2, the normalized MSE is also discontinuous and numerical instable in the cases of $\alpha \approx 1$, where $p_\alpha(x;\ \theta) \stackrel{\text{def}}{=} \frac{m_\alpha(x;\tilde{\theta})}{\sum_x m_\alpha(x;\tilde{\theta})}$ and $p_q(x) \stackrel{\text{def}}{=} \frac{\tilde{q}(x)}{\sum_x \tilde{q}(x)}$. The results for $\alpha \leq -1$ and $\alpha = 1$ are the minimum.

Moreover, we have confirmed the same values of m_α and $\tilde{\theta}$ are estimated for the conventional gradient system (8) by 4th-order Runge-Kutta approximation with $\Delta t = 10^{-4}$, though it has more than 30,000 iterations until the convergence within the same error level as shown in Table 2.

Now, we will discuss the reason why the value of m_α is extremely increased (it causes numerical overflow of MSE), as the value of α approaches to 1 ± 0.

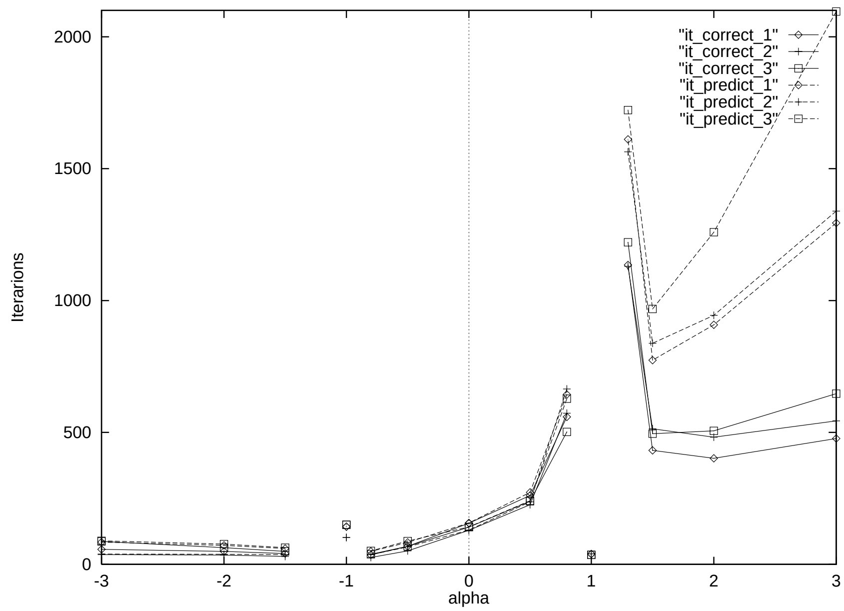

Figure 3: Iteration times for the different data 1, 2, 3 (Solid line: predictor-corrector method, Dashed line: piecewise-linear method).

α	-3	-2	-1.5	-1	-0.8	-0.5	0	0.5	0.8	1	1.3	1.5	2	3
data 1	13	13	13	13	14	14	17	30	107	13	81	39	20	15
data 2	7	7	7	8	8	9	11	25	96	6	74	30	12	8
data 3	6	6	6	6	6	7	9	18	76	6	80	34	14	8

Table 2: Normalized MSE $\frac{1}{m}\sum_x (p_\alpha(x;\, \theta) - p_q(x))^2 \times 10^{-6}$.

Form (1), we consider the function value

$$m_\alpha(x;\, \tilde{\theta}) = \left(\pm\frac{1-\alpha}{2}\right)^{\frac{2}{1-\alpha}} \times \left[\mp\sum_i \tilde{\theta}^i F_i(x)\right]^{\frac{2}{1-\alpha}} = \gamma^{\pm\frac{1}{\gamma}} \times [\sigma_x]^{\pm\frac{1}{\gamma}} \tag{15}$$

where we denote $\sigma_x \stackrel{\text{def}}{=} \mp\sum_i \tilde{\theta}^i F_i(x) > 0$ $(x = 1, 2, \dots, m)$ and $\gamma \stackrel{\text{def}}{=} \pm\frac{1-\alpha}{2} > 0$, the sign $\pm$ are corresponded to $\alpha < 1$ and $\alpha > 1$, respectively.

As $\alpha \to 1 \pm 0$ (equivalently $\gamma \to 0$), the first term in the right-hand side of (15) quickly approaches to 0 ($\alpha < 1$) or ∞ ($\alpha > 1$). However, it should be canceled out in keeping finite range for the updated value of m_α to minimize the divergence D_α. If the first term is dominant, $m_\alpha(x;\, \tilde{\theta}) \to 0$ or ∞ is obtained for all of $x \in \mathcal{X}$, obviously, it is not the desirable estimation value for a finite non-zero data $\tilde{q}(x)$. Therefore, the second term needs to be extremely large ($\alpha < 1$) or small ($\alpha > 1$). By the asymptotic characteristic of the graph $y = [\sigma_x]^{\pm\frac{1}{\gamma}}$, $\sigma_x \to \infty$ is deduced in both cases of $\alpha < 1$ and $\alpha > 1$. Consequently, the estimated value $\pm\tilde{\theta}^i$ is extremely large at least one element, and it causes the numerical overflow. This is a common estimation property independent of data, in the families of the power-functions.

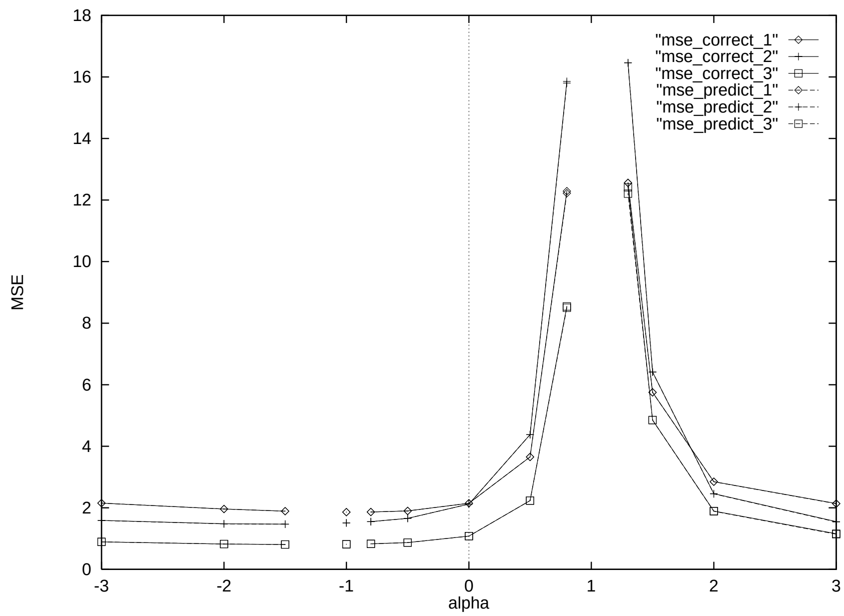

Figure 4: MSE $\frac{1}{m}\sum_x (m_\alpha(x;\, \tilde{\theta}) - \tilde{q}(x))^2$ for the different data 1, 2, 3.

6 As a subclass of convex programming problems

Let us reconsider the α-projection problem as a CP problem. For the SDP problem [14], we can directly consider the self-concordant barrier function $\phi(z)$. However, the corresponded α-divergence in Table 1 is generally not a barrier function. Thus, we reformulate the α-projection problem to a l_p-programming and the related ones, and derive self-concordant barrier functions [8]. Furthermore, we discuss a relation between the $(-\alpha)$-geodesic and the AS-trajectory for a modified barrier function [8]. Although the existence of a self-concordant barrier function for any convex region has been proved, it is generally difficult to explicitly construct it [13].

6.1 Fundamental property of self-concordance

Before beginning of the discussion, we explain the fundamental property of self-concordant barrier function. It belongs to a special subclass of barrier functions with the following smoothness condition [9]:

> Let $\mathcal{F}^0$ be an open convex subset of R^n. A function $\phi(z)$: $\mathcal{F}^0 \to R$ is called κ-self-concordant on $\mathcal{F}^0$, $\kappa > 0$, if $\phi(z)$ is three times continuously differentiable in $\mathcal{F}^0$ and if for all $z \in \mathcal{F}^0$ and $v \in R^n$ the following inequality holds:
>
> $$\nabla^3\phi(z)[v,v,v] \leq 2\kappa(v^T\nabla^2\phi(z)v)^{\frac{3}{2}},$$
>
> where $\nabla^3\phi(z)[v,v,v]$ denotes the third differential of $\phi(z)$ at z and $v^{\ddagger}$.

Self-concordance is considered as a natural extension of logarithmic barrier functions for polyhedral or quadratic constraints, to guarantee a nice performance of Newton's method for CP problems. The most important point is that the gap between the current object function value and the optimal is reduced by a factor in a number of iterations, which is bounded by a polynomial in the problem-size and the self-concordant parameter κ. By using this property, many efficient interior point methods have been developed [9] [10] [13] [16]. They have polynomial rate of convergence for the trajectory through near the central path.

‡ There is a slightly different definition, however the existence of polynomial time algorithms is the same [10] [16].

6.2 Reformulation of the parameter estimation problem

First, we reduce the minimization of the α-divergence for a given data $\tilde{q}(x)$ in the ambient manifold $\tilde{M}$ to an equivalent problem on the submanifold $\tilde{S}_\alpha$.

By the extended Pythagorean theorem [1] [2],

$$D_\alpha(m_\alpha(x;\ \tilde{\theta})||\tilde{q}(x)) = D_\alpha(m_\alpha(x;\ \tilde{\theta})||m_\alpha(x;\ \tilde{\theta}(T))) + D_\alpha(m_\alpha(x;\ \tilde{\theta}(T))||\tilde{q}(x))),$$

holds. Since the 2nd-term in the right-hand side is constant, the minimization of $D_\alpha(m_\alpha(x;\ \tilde{\theta})||\tilde{q}(x))$ is equivalent to the minimization of $D_\alpha(\tilde{\theta})$ denoted by the 1st-term in the right-hand side. The quantity $D_\alpha(\tilde{\theta})$ on $\tilde{S}_\alpha$ is represented as

$$D_\alpha(\tilde{\theta}) = \psi(\tilde{\theta}) - \psi(\tilde{\theta}(T)) + \sum_{i=1}^{n}(\tilde{\theta}^i(T) - \tilde{\theta}^i)\tilde{\eta}_i(T). \tag{16}$$

We should remark $\tilde{\theta}(T)$ and $\psi(\tilde{\theta}(T))$ are not variables but constant values (which depend on only a given data $\tilde{q}(x)$); however, they are unknown in advance before solving the problem. Since the relations $\partial_j D_\alpha(m_\alpha(x;\ \tilde{\theta})||\tilde{q}(x)) = \partial_j D_\alpha(\tilde{\theta})$ and $\partial_i\partial_j D_\alpha(m_\alpha(x;\ \tilde{\theta})||\tilde{q}(x)) = \partial_i\partial_j D_\alpha(\tilde{\theta})$ hold from $\partial_j\psi = \tilde{\eta}_j$ and $\partial_j D_\alpha = \tilde{\eta}_j - \tilde{\eta}_j(T)$, they are not distinguished hereafter. Thus, the difference between $D_\alpha(\tilde{\theta})$ and $D_\alpha(m_\alpha(x;\ \tilde{\theta})||\tilde{q}(x))$ measured by $\tilde{q}(x)$ and the current $\tilde{\theta}$ is buffered in the following variable $\tilde{\theta}^{n+1}$ including the quantities with respect to the unknown $\tilde{\theta}(T)$.

Next, by introducing a slack variable $\tilde{\theta}^{n+1}$, the minimization of $D_\alpha(\tilde{\theta})$ is converted to an equivalent CP problem:

$$\begin{aligned} &\min \quad \tilde{\theta}^{n+1}, \\ &s.t. \quad D_\alpha(\tilde{\theta}) - \tilde{\theta}^{n+1} \le 0. \end{aligned} \tag{17}$$

With additional variables τ_x, $(x = 1, 2, \ldots, m)$, we reformulate the problem (17) to

$$\min \quad \mathbf{c}^T\tilde{\theta},$$

$$s.t. \quad \textstyle\sum_x \tau_x - \mathbf{b}^T\tilde{\theta} - d \le 0, \tag{18}$$

$$\rho(z_x) \le \tau_x, \tag{19}$$

where we denote $\mathbf{b} = (b_1, \ldots, b_n, b_{n+1})^T$, $b_i \stackrel{\text{def}}{=} \tilde{\eta}_i(T)$, $b_{n+1} \stackrel{\text{def}}{=} 1$, $\mathbf{c} \stackrel{\text{def}}{=} (\underbrace{0, \ldots, 0}_{n}, 1)^T$, $d \stackrel{\text{def}}{=} \psi(\tilde{\theta}(T)) - \sum_i \tilde{\theta}^i(T)\tilde{\eta}_i(T)$,

$$\rho(z_x) \stackrel{\text{def}}{=} \pm(z_x)^{\frac{1}{\beta}}, \quad \beta \stackrel{\text{def}}{=} \frac{1-\alpha}{2}, \tag{20}$$

$$z_x \stackrel{\text{def}}{=} \left(\frac{\pm 2}{1+\alpha}\right)^\beta \left[\beta \sum_{i=1}^{n} F_i(x)\tilde{\theta}^i\right], \tag{21}$$

the corresponded sign $\pm$ depends on that of $\frac{2}{1+\alpha}$ to be positive in the parenthesis. Note that (21) is an affine transformation from $\tilde{\theta}$ to a variable z_x. In the above reformulation, we have applied (16) and

$$\psi(\tilde{\theta}) = \frac{2}{1+\alpha}\sum_{x=1}^{m} m_\alpha(x;\ \tilde{\theta}) = \sum_{x=1}^{m}\rho(z_x),$$

to the constraint $D_\alpha(\tilde{\theta}) - \tilde{\theta}^{n+1} \le 0$. Here, at the optimal $\tilde{\theta} \to \tilde{\theta}(T)$, $D_\alpha(\tilde{\theta}) \to 0$ holds form (16), then $\tilde{\theta}^{n+1} \to 0$.

Let us consider self-concordant barrier functions for the reformulated problem. For the constraint (18), we have a self-concordant barrier function (as a logarithmic barrier function)

$$-\ln(d - \sum_x \tau_x + \mathbf{b}^T\tilde{\theta}),$$

case		self-concordant barrier
$-1 < \alpha < 1$	$(0 < \beta < 1,\, p > 1)$	$-\ln(\tau_x^\beta - z_x) - \ln \tau_x$
$\alpha \geq 3$	$(\beta \leq -1,\, -1 \leq p < 0)$	$-\ln(\tau_x - z_x^p) - \ln z_x$
$1 < \alpha < 3$	$(-1 < \beta < 0,\, p < -1)$	$-\ln(z_x - \tau_x^\beta) - \ln \tau_x$
$\alpha < -1$	$(\beta > 1,\, 0 < p < 1)$	$-\ln(z_x^p + \tau_x) - \ln z_x$
$\alpha = -1$	(mixture family)	$-\ln(\tau_x - z_x(\ln z_x - 1)) - \ln z_x$

Table 3: Self-concordant barrier functions for the the constraint (19).

because of the linear inequality [10] [13] for the $m+n+1$ variables of τ and $\tilde{\theta}$. This function is common for any values of α.

In addition, for the constraint (19), independently considering each epigraph of a power-function for the two dimensional set (z_x, τ_x), we have derived $\frac{5}{3}$-self-concordant barrier functions [9] according to the values of α as shown in Table. 3. For the constraints (18) and (19), we have the following $\left(1 + \frac{5m}{3}\right)$-self-concordant barrier function [8]

$$-\ln(d - \sum_x \tau_x + \mathbf{b}^T\tilde{\theta}) - \sum_x \left\{ \ln(\tau_x^\beta - z_x) + \ln \tau_x \right\}, \tag{22}$$

where the 2nd term is replaced by them in Table 3 according to the values of α.

6.3 Relation between geodesic and affine scaling trajectory

Let us consider a barrier function

$$h(\tilde{\theta}) \stackrel{\text{def}}{=} -\ln(\ \delta(\tilde{\theta})\ ), \quad \delta(\tilde{\theta}) \stackrel{\text{def}}{=} \tilde{\theta}^{n+1} - D_\alpha(\tilde{\theta}). \tag{23}$$

If $h(\tilde{\theta})$ is self-concordant, the problem (17) is solvable in polynomial time without the reformulation of (20) and (21). However, this case requires the special conditions [13] as shown in subsection 6.1 for the 2nd and 3rd derivatives of D_α.

Otherwise, in the following discussion, it may approximate the AS-trajectory for the self-concordant barrier function (22) near the boundary $\tau_x \approx \rho(z_x)$.

By the components of partial derivatives

$$\frac{\partial h}{\partial \tilde{\theta}^i} = \frac{\partial_i D_\alpha}{\delta}, \quad \frac{\partial h}{\partial \tilde{\theta}^{n+1}} = -\frac{1}{\delta},$$
$$\frac{\partial^2 h}{\partial \tilde{\theta}^i \partial \tilde{\theta}^j} = \frac{\partial_i \partial_j D_\alpha \times \delta + \partial_i D_\alpha \partial_j D_\alpha}{\delta^2},$$
$$\frac{\partial^2 h}{\partial \tilde{\theta}^i \partial \tilde{\theta}^{n+1}} = -\frac{\partial_i D_\alpha}{\delta^2},$$
$$\frac{\partial^2 h}{\partial^2 \tilde{\theta}^{n+1}} = \frac{1}{\delta^2},$$

$i, j \in \{1, 2, \ldots, n\}$, the Hessian of h is obtained as

$$\frac{1}{\delta^2} \begin{bmatrix} [\partial_i \partial_j D_\alpha] \times \delta + (\partial_i D_\alpha)(\partial_j D_\alpha)^T & (-\partial_i D_\alpha) \\ (-\partial_i D_\alpha)^T & 1 \end{bmatrix} \stackrel{\text{def}}{=} \frac{1}{\delta^2} \begin{bmatrix} A & \mathbf{B} \\ \mathbf{B}^T & 1 \end{bmatrix},$$

where A is a $n \times n$ matrix, $\mathbf{B}$ is a n-vector.

From the inverse matrix formula, we obtain

$$\begin{bmatrix} A & \mathbf{B} \\ \mathbf{B}^T & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \Pi^{-1} & -\Pi^{-1}\mathbf{B} \\ -\mathbf{B}^T\Pi^{-1} & 1+\mathbf{B}^T\Pi^{-1}\mathbf{B} \end{bmatrix}, \tag{24}$$

$$\Pi \stackrel{\text{def}}{=} A - \mathbf{B}\mathbf{B}^T = \delta \times [\partial_i \partial_j D_\alpha],$$

$$\Pi^{-1} = \frac{1}{\delta} [\partial_i \partial_j D_\alpha]^{-1}.$$

Note that the Newton's direction $-[\nabla^2 h]^{-1}\nabla h$ for the barrier function is given from (24) as follows:

$$\begin{pmatrix} d\tilde{\theta}/dt' \\ d\tilde{\theta}^{n+1}/dt' \end{pmatrix} = \delta \times \begin{bmatrix} A & \mathbf{B} \\ \mathbf{B}^T & 1 \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{B} \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \delta \end{pmatrix}.$$

By $\delta > 0$ in (23), the analytic center [9] is defined as $\tilde{\theta}^{n+1} \to \infty$ for any values of $(\tilde{\theta}^1, \ldots, \tilde{\theta}^n)$. We should set only the initial value of $\tilde{\theta}^{n+1}$ as large as possible.

On the other hand, the direction of AS-trajectory $-[\nabla^2 h]^{-1}\mathbf{c}$ is also given from (24) as follows:

$$\frac{d\tilde{\theta}^i}{dt'} = \delta^2 \times \Pi^{-1}\mathbf{B} = -\delta \sum_j [\partial_i \partial_j D_\alpha]^{-1} \partial_j D_\alpha, \tag{25}$$

$$\frac{d\tilde{\theta}^{n+1}}{dt'} = -\delta^2(1 + \mathbf{B}^T\Pi^{-1}\mathbf{B}) = -\delta(\delta + \|\Delta D_\alpha\|^2_{G^{-1}}) \;\leq\; 0,$$

where $\|\Delta D_\alpha\|^2_{G^{-1}} = \|\Delta\tilde{\theta}\|^2_G = \sum_{ij} \partial_i D_\alpha [\partial_i \partial_j D_\alpha]^{-1} \partial_j D_\alpha$ is obtained from the difference (discrete) version of (8). Remember that $[g^{ij}] = [g_{ij}]^{-1} = [\partial_i \partial_j D_\alpha]^{-1}$.

Therefore, the direction vector (25) is the same to the gradient (8) for the $(-\alpha)$-geodesic, and the magnitude is multiplied by $\delta(\tilde{\theta})$ as the difference between $D_\alpha(\tilde{\theta})$ and an upper-bound value $\tilde{\theta}^{n+1}$. These results hold on more general ∇^*-geodesic for minimizing a ∇-divergence [1] [2]. From the relation (23) between $h(\tilde{\theta})$ and $D_\alpha(\tilde{\theta})$, the AS-trajectory (25) for $h(\tilde{\theta})$ and the gradient system (8) as continuous Newton's method for $D_\alpha(\tilde{\theta})$ have been connected. They move on the same ∇^*-geodesic with different speeds by $\delta(\tilde{\theta})$.

7 Conclusion

To develop efficient algorithms such as interior-point methods for CP problems [9] [13], we have studied a common structure in a class of dynamical system related to Newton's method. On the information geometric structure [1] [2], we have proposed efficient calculation methods [7] for the estimation problem onto a space of power-functions. For the α-projection to minimize the α-divergence as a convex functional, the maximum step-length is derived in the piece-wise linear approximation and predictor-corrector methods. They are similar to the approximations of an AS-trajectory for SDP problems [14] based on the autoparallelism, in spite of the quite different formulations. By simulation results, it has been shown that the proposed methods have more less iterations compared to the conventional Runge-Kutta approximation. We have also discussed the estimation property according to the value of α. Furthermore, the α-projection problem has been reconsidered [8] as a form of CP problem. From the reformulated l_p-like problem, we have derived self-concordant barrier functions, and pointed out the coincidence with the gradient directions on the geodesic for the divergence and on the AS-trajectory for a modified barrier function. These results give new geometric and algorithmic insights into nonlinear analysis of dynamical systems by Newton's methods for functional estimation.

References

[1] S. Amari, Differential-geometrical methods in statistics, Springer Lecture Notes in Statistics, vol. 28, 1985.

[2] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Oxford Univ. Press, Translations of Mathematical Monographs, Vol.191, 2000.

[3] H. Dyckhoff and W. Pedrycz, Generalized means as model of compensative connectives, *Fuzzy Sets and Systems*, Vol. 14, pp. 143-154, 1984.

[4] A. Fujiwara and S. Amari, Gradient systems in view of information geometry, *Physica D*, Vol. 80, pp. 317-327, 1995.

[5] G. H. Golub and C. F. Van Loan, *Matrix Computations,* 3rd ed., Johns Hopkins U.P., 1996.

[6] Y. Hayashi, Direct Calculation Methods for Parameter Estimation in Statistical Manifolds of Finite Discrete Distributions, *IEICE Trans. on Fundamentals*, Vol. E81-A, No. 7, pp. 1486-1492, July 1998.

[7] Y. Hayashi, Adaptive control of averaging operators on geodesics with the enhancement of gradient descent or ascent, *Proc. of International Symposium on NOLTA*, Vol. 2, pp. 667-670, 1999.

[8] Y. Hayashi, A New Relation between Information Geometry and Convex Programming - Coincidence with the Gradient Vectors for the Divergence and a Modified Barrier Function, *To appear in IEICE Fundamentals*, 2001.

[9] D. den Hertog, *Interior Point Approach to Linear, Quadratic and Convex Programming*, Kluwer Academic Publishers, 1994.

[10] F. Jarre, Interior-point methods for classes of convex programs, in Interior Point Methods of Mathematical Programming, ed. T. Terlaky, pp. 255-296, Kluwer Academic Publishers, 1996.

[11] W. Menke, *Geophysical Data Analysis: Discrete Inverse Theory,* Revised Edition, Academic Press, 1989.

[12] M. Mizumoto, Pictorial representations of fuzzy connectives, Part I: Cases of t-norms, t-conorms and averaging operators, *Fuzzy Sets and Systems*, Vol. 31, pp. 217-242, 1989.

[13] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Studies in Applied Mathematics, 1994.

[14] A. Ohara, *Information Geometric Analysis of a Interior-Point Method for Semidefinite Programming*, RIMS Kokyuroku, pp. 71-89, 1997.

[15] E. Polak, Optimization: Algorithms and consistent approximations, Springer-Verlag New York, Applied Mathematical Sciences, Vol. 124, 1997.

[16] T. Tsuchiya, New progress of optimization algorithms - Interior point method and the peripherals V, System/Control/Information, vol. 42, no. 12, pp 677-686, 1999 (in Japanese).

A Appendix: Maximum step-length in the cases of $\alpha = \pm 1$

As similar to the case of α-affine manifold, we derive the maximum step-length δt for the exponential and mixture families. In these cases, the estimation problems are based on the minimization of Kullback-Leibler divergence and the dual one [1] [2].

Let us consider the following functions p_E and p_M instead of m_α

Exponential family:

$$p_E(x;\theta) \stackrel{\text{def}}{=} \exp\left[\sum_i F_i(x)\theta^i - \psi(\theta)\right] = \frac{\exp\left[\sum_i F_i(x)\theta^i\right]}{\sum_x \exp\left[\sum_i F_i(x)\theta^i\right]}.$$

Mixture family:

$$p_M(x;\tilde{\theta}) \stackrel{\text{def}}{=} \sum_i F_i(x)\tilde{\theta}^i.$$

Here, (2) and (13) hold in both cases. Note that the exponential family is an exception[§] in the families of $\tilde{S}_\alpha$, in the sense it has the condition of normalization.

A.1 Exponential family ($\alpha = 1$)

Since the m-geodesic is represented as a straight line of the dual parameter $\eta_j = \sum_x F_j(x)p_E(x;\theta)$, the linearity of p_E is considered in the Talyor expansion by the time-variable t. The first and second derivatives are as follows:

$$\frac{dp_E}{dt} = \sum_i \frac{\partial p_E}{\partial \theta^i}\frac{d\theta^i}{dt} = p_E \sum_i (F_i - \eta_i)\frac{d\theta^i}{dt},$$

$$\frac{d^2 p_E}{dt^2} = p_E \sum_{i,j}\left\{(F_i - \eta_i)(F_j - \eta_j) - g_{ij} + \sum_k (F_k - \eta_k)\Gamma_{ij}^k\right\}\frac{d\theta^i}{dt}\frac{d\theta^j}{dt},$$

where the Fisher information metric and the m-connection coefficient are

$$g_{ij} = \sum_x (F_i(x) - \eta_i)(F_j(x) - \eta_j)p_E(x;\theta),$$

$$\Gamma^*_{ijh} = \sum_x (F_h(x) - \eta_h)\left\{(F_i(x) - \eta_i)(F_j(x) - \eta_j) - g_{ij}\right\} p_E(x;\theta).$$

From the condition $\left|\frac{1}{2}\frac{d^2 p_E}{dt^2}\delta t\right| < \varepsilon^2$, we obtain

$$\delta t < \min_x \frac{\varepsilon}{\sqrt{\frac{p_E(x;\theta)}{2}\left|\sum_{i,j}\left\{(F_i(x) - \eta_i)(F_j(x) - \eta_j) - g_{ij} + \sum_k (F_k(x) - \eta_k)\Gamma_{ij}^k\right\}\frac{d\theta^i}{dt}\frac{d\theta^j}{dt}\right|}}.$$

A.2 Mixture family ($\alpha = -1$)

Since the e-geodesic is represented as a straight line of the dual parameter $\tilde{\eta}_j = \sum_x F_j(x)\log p_M(x;\tilde{\theta})$, the linearity of $\log p_M$ is considered in the Talyor expansion by the time-variable t. The first and second derivatives are as follows:

$$\frac{d\log p_M}{dt} = \sum_i \partial_i \log p_M \frac{d\tilde{\theta}^i}{dt} = \frac{1}{p_M}\sum_i F_i \frac{d\tilde{\theta}^i}{dt},$$

[§] As a distinction, the parameters θ and η are denoted without tilde.

$$\frac{d^2 \log p_M}{dt^2} = -\frac{1}{p_M} \sum_{i,j} \left\{ \frac{F_i F_j}{p_M} - \sum_k F_k \Gamma_{ij}^k \right\} \frac{d\tilde{\theta}^i}{dt} \frac{d\tilde{\theta}^j}{dt},$$

where the Fisher information metric and the e-connection coefficient are

$$g_{ij} = \sum_x \frac{F_i(x) F_j(x)}{p_M(x; \tilde{\theta})},$$

$$\Gamma_{ijh}^* = -\sum_x \frac{F_i(x) F_j(x) F_h(x)}{p_M(x; \tilde{\theta})^2}.$$

From the condition $\left| \frac{1}{2} \frac{d^2 \log p_M}{dt^2} \delta t \right| < \varepsilon^2$, we obtain

$$\delta t < \min_x \frac{\varepsilon}{\sqrt{\frac{1}{2 p_M(x;\tilde{\theta})} \left| \sum_{i,j} \left\{ \frac{F_i(x) F_j(x)}{p_M(x;\tilde{\theta})} - \sum_k F_k(x) \Gamma_{ij}^k \right\} \frac{d\tilde{\theta}^i}{dt} \frac{d\tilde{\theta}^j}{dt} \right|}}.$$

Moreover, the corrector process results in systems of linear equations derived form the definitions of p_E and p_M, respectively. They are also solvable by using QR-factorization [6] as similar to the cases of $\alpha \neq \pm 1$.