Adaptive control of averaging operators on geodesics with the enhancement of gradient descent or ascent

JAIST, Yukio HAYASHI (yhayashi@jaist.ac.jp)

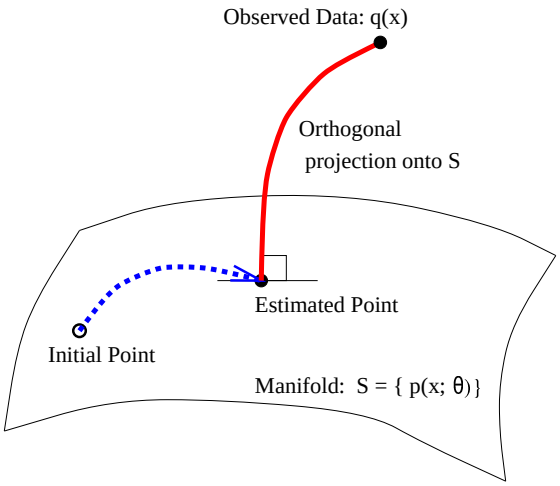
Contents

1. Introduction (efficient learning and optimization algo.)
2. Our Previous Works (in the cases of $\alpha = \pm 1$)
3. Extension to an α-affine manifold
   Efficient caluc. of $(-\alpha)$-geodesic and parameter estimation
4. Estimation Property
   Enhancement of the Gradient
   Estimation Error and Numerical Stability
5. Averaging Operators and Foliation (control of α)
6. Conclusion

—

1. Introduction

1-1. Parameter estimation problem



Find the parameter value to minimize a quasi-distance from a data onto the manifold $\Rightarrow$ computing the geodesic

—

1. Introduction

1-2. Natural gradient method

When the parametric function space is Riemannian, natural gradient method through the geodesic shows good performance.

On-line Learning for MLP (Amari '98): adaptively minimize $l(y|x;\theta) \stackrel{\text{def}}{=} \log p(y|x;\theta)$, and show fast convergence as Newton's method of the batch version:

$$\frac{d\theta^i}{dt} = -\sum_j \left[\frac{\partial^2 D_1}{\partial\theta^i\partial\theta^j}\right]^{-1} \frac{\partial D_1}{\partial\theta^j}, \quad g_{ij} = \frac{\partial^2 D_1}{\partial\theta^i\partial\theta^j} \neq E_q\left[\frac{\partial^2 l(y|x;\theta)}{\partial\theta^i\partial\theta^j}\right]$$

$$\text{on-line version} \Downarrow \nabla D_1 = E_q[\nabla l(y|x;\theta)]$$

$$\boxed{\theta_{t+1} = \theta_t - \varepsilon_t G_t(\theta_t)^{-1}\nabla l(y_t|x_t;\theta_t)}$$

However, $[g_{ij}]^{-1}$ requires much computations at each iteration.
$\Rightarrow$ Approximation of $[g_{ij}]^{-1}$ using Kalman Filter technique for various stochastic models: NN, regression, classification.

—

1. Introduction

1-3. Interior point method and geodesic

min $c^T x$: linear object function

s.t. $x \in \mathcal{M} \subset R^m$: convex region

Let us consider a barrier function $\Psi_t(x) \stackrel{\text{def}}{=} t \times c^T x + \psi(x)$.
From the optimal condition $\partial \Psi_t(x)/\partial x^i = t c_i + \partial \psi(x)/\partial x^i = 0$, we have the solution $y_i(t) = -c_i t$ or $\boxed{dy_i/dt = -c_i}$ in the dual coordinate system $y_i \stackrel{\text{def}}{=} \partial \psi(x)/\partial x^i$.

Affine scaling trajectory $\Downarrow$ $\frac{dy_i}{dt} = \sum_j \frac{\partial y_i}{\partial x^j}\frac{dx^j}{dt} = \sum_j g_{ij}\frac{dx^j}{dt}$

$$\frac{dx^j}{dt} = -\sum_i g^{ji} c_i$$

But the inverse trans. $y(t) \to x(t)$ is generally implicit !!
$\Rightarrow$ Non-iterative IPM derived from autoparallelism in a special case of SDP problems (Ohara '97)

—

2. Our Previous Works (in the cases of $\alpha = \pm 1$)

As similar to the case of SDP, we have proposed efficient calculation methods for em-geodesic defined by the gradient system

$$\frac{d\theta^i}{dt} = -\sum_j g^{ij}(\theta)\frac{\partial D_{\pm 1}(\theta)}{\partial \theta^j}$$

on the submanifold of mixture or exponential family, applying the linearity of $\log p(x;\ \theta)$: Mix or $p(x;\ \theta)$: Exp in the case of em-autoparallel.

$\Rightarrow$ We will extend the result to more general α-affine manifold ($\alpha \neq \pm 1$, $\alpha \in R$) in the function space over finite discrete variables $x \in \mathcal{X} = \{1, 2, \ldots, m\}$.

Note: We focus finite discrete variables with our interest in both theory and application for computer simulation (a kind of discrete inverse problem).

—

3. Extension to an α-affine manifold

3-1. α-affine manifold

Let us consider n-dimensional submanifold $\tilde{S}_\alpha = \{m_\alpha(x;\ \tilde{\theta})\}$.

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

We will show efficient algorithms for

- $(-\alpha)$-geodesic between any given two points $\tilde{\theta}(p)$ and $\tilde{\theta}(q)$.
- Parameter estimation onto the submanifold from an observed data $\tilde{q}(x)$.

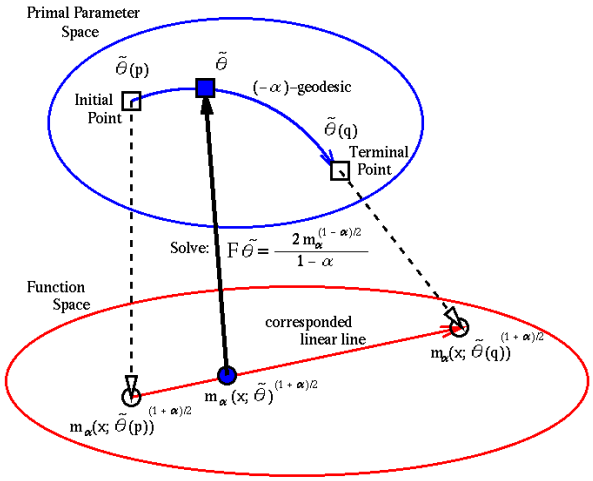
They are solvable through the system of linear equations $\boxed{F\tilde{\theta} = \frac{2}{1-\alpha} m_\alpha^{\frac{1-\alpha}{2}}}$ instead of the iterative gradient system.
As the inverse operation, the QR-factorization of F can be commonly used for any initial, terminal, and the interpolated points.

—

3. Extension to an α-affine manifold

3-2. Direct calculation for $(-\alpha)$-geodesic

the geodesic $\Leftrightarrow$ linear line of $\tilde{\eta}_j \stackrel{\text{def}}{=} \frac{2}{1+\alpha} \sum_j F_j(x) \left[m_\alpha(x;\ \tilde{\theta}) \right]^{\frac{1+\alpha}{2}}$



3. Extension to an α-affine manifold

3-3. Parameter estimation onto the submanifold from an observed data $\tilde{q}(x)$

The linearity of $\left[m_\alpha(x;\ \tilde{\theta})\right]^{\frac{1+\alpha}{2}}$ is also applied for the parameter estimation problem, which is transformed into the following double systems of linear equations:

$$\begin{cases} F^T m_\alpha^{\frac{1+\alpha}{2}} = \frac{1+\alpha}{2}\tilde{\eta}(\tilde{q}), \quad \tilde{\eta}(\tilde{q}) \stackrel{\text{def}}{=} \frac{2}{1+\alpha}F^T\tilde{q}^{\frac{1+\alpha}{2}}, \\ F\tilde{\theta} = \frac{2}{1-\alpha}m_\alpha^{\frac{1-\alpha}{2}} \end{cases}$$

The first equations have many candidates on the projection curve from the data $\tilde{q}(x)$ onto $\tilde{S}_\alpha$, it is not solvable in this form.
$\Rightarrow$ Since the solution gives the minimum α-divergence, a proper line-search algorithm can be applied in the case of autoparallel.

—

3. Extension to an α-affine manifold

3-4. Direct estimation method

Consider a linear bounded time-scale $0 \leq t' \leq 1$, from the solution of the geodesic $\tilde{\eta}_j = \tilde{\eta}_j(p) + \Delta\tilde{\eta}_j \times t'$, $\Delta\tilde{\eta}_j \stackrel{\text{def}}{=} \tilde{\eta}_j(q) - \tilde{\eta}_j(p)$,

$$\frac{d\tilde{\theta}^i}{dt'} = \sum_j \frac{\partial\tilde{\theta}^i}{\partial\tilde{\eta}_j}\frac{d\tilde{\eta}_j}{dt'} = \sum_j g^{ij}(\tilde{\theta})\Delta\tilde{\eta}_j$$

is obtained (However, the meaning of $\min D_\alpha$ disappears).
When the function $m_\alpha^{\frac{1+\alpha}{2}}$ is linear, the estimated function m_α is easily obtained from the following equation at $t' = 1$,

$$m_\alpha(x;\ \tilde{\theta})^{\frac{1+\alpha}{2}} = m_\alpha(x;\ \tilde{\theta}(p))^{\frac{1+\alpha}{2}} + \Delta m_\alpha(x)^{\frac{1+\alpha}{2}} \times t'.$$

Here, the constant direction

$$\Delta m_\alpha(x)^{\frac{1+\alpha}{2}} = m_\alpha(x;\ \tilde{\theta} + \Delta\tilde{\theta})^{\frac{1+\alpha}{2}} - m_\alpha(x;\ \tilde{\theta})^{\frac{1+\alpha}{2}}$$

is calculated by the only one-step update of the gradient system.
Thus, there are no accumulated error in the estimated function.

—

4. Estimation Property

4-1. Enhancement of the gradient

When we introduce an extension of log functional

$$L^{(\alpha)}(z) \stackrel{\text{def}}{=} \frac{2}{1+\alpha}\left\{z^{\frac{1+\alpha}{2}} - 1\right\}, \quad (\alpha \neq -1)$$

we can rewrite

$$D_\alpha = \frac{2}{1-\alpha}\sum_x \left[m_\alpha \left\{L^{(1)}\left(\frac{\tilde{q}}{m_\alpha}\right) - L^{(\alpha)}\left(\frac{\tilde{q}}{m_\alpha}\right)\right\}\right],$$

$$-\partial_j D_\alpha = \sum_x \left\{L^{(\alpha)}\left(\frac{\tilde{q}}{m_\alpha}\right) \times \partial_j m_\alpha\right\}, \quad \partial_j \stackrel{\text{def}}{=} \frac{\partial}{\partial \tilde{\theta}^j}$$

Thus, $L^{(\alpha)}$ is regarded as a function weight for the gradient $\partial_j m_\alpha$ at each variable $x \in \mathcal{X} = \{1, 2, \ldots, m\}$.

—

4. Estimation Property

4-2. Effect of the function weight

Gradient descent or ascent according to the relation of functions $\tilde{q}(x)$ and $m_\alpha(x;\ \tilde{\theta})$ in the inequality (negative or positive weight). In both cases of grad. asc. and des., it is expected $m_\alpha \longrightarrow \tilde{q}$.

Grad.	Descent: $-\partial_j m_\alpha$	Ascent: $\partial_j m_\alpha$
Func.	$\tilde{q} < m_\alpha$	$\tilde{q} > m_\alpha$
$\alpha < 1$	$L^{(\alpha)} < L^{(1)} < 0$	$0 < L^{(\alpha)} < L^{(1)}$
$\alpha > 1$	$L^{(1)} < L^{(\alpha)} < 0$	$0 < L^{(1)} < L^{(\alpha)}$

As $|L^{(\alpha)}|$ is larger, the gradient is enhanced in comparison with that at $\alpha \approx 1$. However there is a trade-off between the effects of $-\partial_j D_\alpha$ and g^{ij} in the gradient system $d\tilde{\theta}^i/dt = -\sum_j g^{ij}\partial_j D_\alpha$

As $\alpha \to \pm\infty$, $|L^{(\alpha)}| \to$ larger, but $|g^{ij}| \to$ smaller.

—

4. Estimation Property

4-3. Numerical simulation of the gradient system

We have investigated the total effect of the function weights through numerical simulation ($n = 9$, $m = 125$)

	$\alpha < -1$	mix: $\alpha = -1$	$-1 < \alpha < 0$	$0 \leq \alpha < 1$	exp: $\alpha = 1$	$\alpha > 1$
MSE	almost const.	quick conv.	as $\alpha \to 0$ slightly up	at $\alpha \approx 1$ extremely growth $\nearrow \infty$	reasonable cov. (discont.)	from $\alpha \approx 1$ settle down $\infty \searrow$
θ	enlarged as α decrease	contrast among θ	continuously changed	at $\alpha \approx 1$ overflow $\nearrow \infty$	weak contrast	from $\infty \searrow$ $\theta^i \approx 0$ as α increase
p_α	slightly smaller fluct.	over estimate for $\alpha = 1$	slightly larger fluct.	at $\alpha \approx 1$ overflow $\nearrow \infty$	reasonable estimate	from $\infty \searrow$ become flat

$\Rightarrow$ Numerical instable with very slowly convergence at $\alpha \approx 1$

—

5. Averaging Operators and Foliation

5-1. Fuzzy averaging operators

Let us consider the function m_α regarded as a generalized mean (called CES-function: $\beta < 1$, Constant Elasticity Substitution)

$$0 < m_\alpha = \left[\frac{1}{n}\sum_{i=1}^{n} u_i(x)^\beta\right]^{\frac{1}{\beta}} < 1,$$

where $0 < u_i(x) \stackrel{\text{def}}{=} \left[\tilde{\xi}^i F_i(x)\right]^{\frac{1}{\beta}} < 1$, $\tilde{\xi}^i \stackrel{\text{def}}{=} \beta n \tilde{\theta}^i \geq 0$, $\beta \stackrel{\text{def}}{=} \frac{1-\alpha}{2}$.

α	$-\infty$	$<$	-1	$<$	1	∞
β	∞	$>$	1	$>$	0	$-\infty$
mean func.	OR max	OR-like $\leftarrow$	arithmetic (center)	AND-like $\rightarrow$	geometric	AND min

To avoid numerical overflow for u^β and $m_\alpha^{1-\beta} = m_\alpha^{\frac{1+\alpha}{2}}$, $\beta > 0$ and $\beta < 1$ ($-1 < \alpha < 1$) are reasonable.

—

5. Averaging Operators and Foliation

5-2. Foliation of $m_\alpha^{\frac{1+\alpha}{2}}$ in the function space $\tilde{M}$

At any points in the coordinate system of $u(x)$, the following monotone relations hold for $-1 < \alpha < \alpha' < 1$,

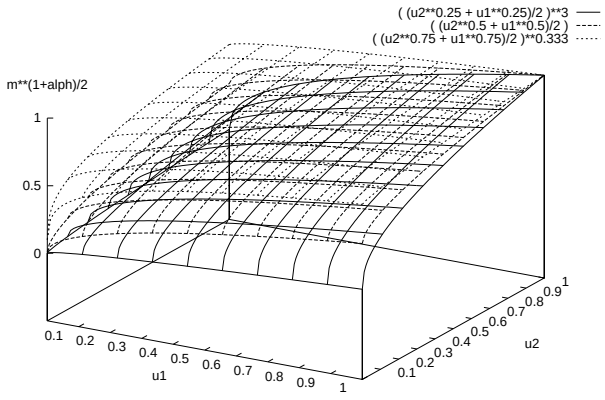
$$m_{\alpha'} \leq m_\alpha, \qquad m_{\alpha'}^{\frac{1+\alpha'}{2}} < m_\alpha^{\frac{1+\alpha}{2}}.$$

Equality holds when $u_1(x) = u_2(x) = \ldots = u_n(x) = u$, which means a hyper-line in the function space of m_α as the one parameter extended family of $\tilde{S}_\alpha$.
There are asymptotical two diagonal points $m_\alpha^{\frac{1+\alpha}{2}} \to 0, 1$, as $u \to 0, 1$ for all values of α.

—

5. Averaging Operators and Foliation

5-3. 2-dimensional case for u_1 and u_2



By control of α, the trajectory traverses in the foliation between arithmetic and geometric means. (AND-like: $-1 < \alpha < 1$).

—

6. Conclusion

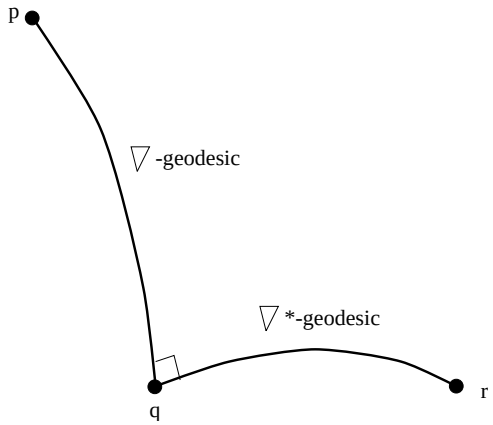
For the geodesic and the parameter estimation ($\min D_\alpha$) on an α-affine manifold over finite discrete variables,

- We have proposed efficient calculation methods without any iterations by using the (invertible) linearity in the case of autoparallel, which may give a potentialy good mathematical structure for learning or optimization method.
- Introducing an extension of logarithm, we have discussed the enhancement of the gradient. The total effect has been investigated by estimation error and numerical stability.
- Moreover, we have pointed out the logical (fuzzy averaging operator) and geometrical (foliation) meanings.

$\Rightarrow$ The estimation property is clarified in the function class.

—

Appendix 1. Extended Pythagorean Theorem



For the ∇-divergence,

$$D(p||r) = D(p||q) + D(q||r)$$

is satisfied on a dually-flat manifold at triangular points p, q, r.

—

Appendix 2. What is natural to minimize $D_\alpha(\theta)$?

We derive the steepest direction $\mathbf{d}\theta = \varepsilon\mathbf{a}$ to minimize $D_\alpha(\theta)$ in a Riemannian space, under the constraint

$$||\mathbf{a}||_G = \sum_{ij} g_{ij} a_i a_j = 1.$$

The search direction vector $\mathbf{a}$ minimizes

$$D_\alpha(\theta + \mathbf{d}\theta) = D_\alpha(\theta) + \varepsilon \nabla D_\alpha(\theta)^T \mathbf{a}.$$

By the Lagrangean method, we have

$$\frac{\partial}{\partial a_j} \left\{ \nabla D_\alpha(\theta)^T \mathbf{a} - \lambda \mathbf{a}^T G \mathbf{a} \right\} = 0.$$

From this euqation, the metric G^{-1} naturally arrears in

$$\mathbf{a} = \frac{1}{2\lambda} G^{-1} \nabla D_\alpha(\theta),$$

where λ is determined from the constraint.

—

Appendix 3. Weighted EM (Matsuyama '97)

In the M-step, we minimize the α-divergence

$$D_\alpha(p_i||q_i) = \frac{4}{1-\alpha^2}\left\{1 - \sum_i p_i \left(\frac{q_i}{p_i}\right)^{\frac{1+\alpha}{2}}\right\} = \frac{-2}{1-\alpha}\sum_i p_i L^{(\alpha)}\left(\frac{q_i}{p_i}\right).$$

By the maximization of $L^{(\alpha)}(p_{Y|X,\phi})$ for a NN, instead of the standard log-likelihood (KL-divergence), the connection weight is updated with the conditional probability weight: $p^{\frac{1+\alpha}{2}}$ $(0 < p < 1, \alpha < -1)$,

$$\Delta U_{ij}^{(\alpha)(\phi)} = p_{Y|X,\phi}(y|x,\phi)^{\frac{1+\alpha}{2}} \times \Delta U_{ij}^{(-1)}(\phi),$$

where $\Delta U_{ij}^{(-1)}(\phi)$ is the direction to maximize the log-likelihood. A large α emphasizes learning at high probability density; however, smaller α's watch errors rigorously at low density parts. Here, x and y denote the complete data contains with unknown parts and the observed data, respectively.

—

Appendix 4. Mixture or exponential family

Let us consider a submanifold $S = \{p(x;\ \theta)\}$ over the finite discrete variables $x \in \mathcal{X} = \{1, 2, \ldots, m\}$.
Mix ($\alpha = -1$):

$$p(x;\theta) \stackrel{\text{def}}{=} C(x) + \sum_{i=1}^{n} \theta^i F_i(x),$$

Exp ($\alpha = +1$):

$$p(x;\theta) \stackrel{\text{def}}{=} \frac{\exp\{C(x) + \sum_{i=1}^{n} \theta^i F_i(x)\}}{\sum_{x\in\mathcal{X}} \exp\{C(x) + \theta^i F_i(x)\}},$$

In our simulation, for the sake of simplicity, we have chosen $C(x) = 0$.

—

Appendix 5. Algorithm for $(-\alpha)$-geodesic

- Compute $m_\alpha(x;\ \tilde{\theta}(p))$ and $m_\alpha(x;\ \tilde{\theta}(q))$ for given two points $\theta(p)$ and $\theta(q)$.
- Select an interpolated point with a rate $0 < \gamma < 1$

$$m_\alpha(x;\ \tilde{\theta})^{\frac{1+\alpha}{2}} = \gamma m_\alpha(x;\ \tilde{\theta}(p))^{\frac{1+\alpha}{2}} + (1-\gamma) m_\alpha(x;\ \tilde{\theta}(q))^{\frac{1+\alpha}{2}}.$$

- Solve the system of linear equation $F\tilde{\theta} = \frac{2}{1-\alpha} m_\alpha^{\frac{1-\alpha}{2}}$

Note: The geodesic is also represented by a linear line of $\tilde{\eta}$:

$$\frac{d\tilde{\eta}_j}{dt} = -(\tilde{\eta}_j - \tilde{\eta}_j(q)).$$

The solution has exponential time-decay:

$$\tilde{\eta}_j = \tilde{\eta}_j(q) + (\tilde{\eta}_j(p) - \tilde{\eta}_j(q))e^{-t}.$$

The one-to-one corresponded $\tilde{\theta}$ and $\tilde{\eta}$ are implicitly connected by the Legendre transformation $\varphi(\tilde{\eta}) = \sum_i \tilde{\theta}^i \tilde{\eta}_i - \psi(\tilde{\theta})$.

—

Appendix 6. Trade-off between the effects of $-\partial_j D_\alpha$ and g^{ij} in the gradient system $\frac{d\tilde{\theta}}{dt} = -\sum_j g^{ij}\partial_j D_\alpha$

$$-\partial_j D_\alpha = \sum_x \left\{ L^{(\alpha)}\left(\frac{\tilde{q}}{m_\alpha}\right) \times \partial_j m_\alpha \right\}.$$

Here, the metric

$$g_{ij} = \sum_x F_i(x) F_j(x) \left[m_\alpha(x;\, \tilde{\theta}) \right]^\alpha$$

is larger at $m_\alpha < 1$ (as $\alpha \to -\infty$) or $m_\alpha > 1$ (as $\alpha \to \infty$).

This table shows the comparison with the effects of $\partial_j m_1$ and $\partial_j m_\alpha$ as the function weights at the condition $m_1 = m_\alpha$

	$\tilde{q} < m_\alpha < 1$	$\tilde{q} > m_\alpha > 1$
$\alpha < 1$	$\partial_j m_1 < \partial_j m_\alpha$ enhancement of grad. des.	$\partial_j m_1 > \partial_j m_\alpha$
$\alpha > 1$	$\partial_j m_1 > \partial_j m_\alpha$	$\partial_j m_1 < \partial_j m_\alpha$ enhancement of grad. asc.

—