Entropy **2014**, *16*, 4338-4352; doi:10.3390/e16084338

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

A Natural Gradient Algorithm for Stochastic Distribution Systems

Zhenning Zhang [1], Huafei Sun [2,*], Linyu Peng [3] and Lin Jiu [4]

[1] Department of Mathematics, Beijing University of Technology, Beijing 100124, China;
E-Mail: zhangzhenning@bjut.edu.cn

[2] Department of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081, China

[3] Department of Applied Mechanics and Aerospace Engineering & Research Institute of Nonlinear PDEs, Waseda University, Okubo, Shinjuku, Tokyo 169-8555, Japan;
E-Mail: l.peng@aoni.waseda.jp

[4] Department of Mathematics, Tulane University, 6823 St. Charles Ave., New Orleans, LA 70118, USA; E-Mail: ljiu@tulane.edu

* Author to whom correspondence should be addressed; E-Mail: huafeisun@bit.edu.cn;
Tel.: +86-10-8257-0539.

*Received: 10 September 2013; in revised form: 15 July 2014 / Accepted: 28 July 2014 /
Published: 4 August 2014*

Abstract: In this paper, we propose a steepest descent algorithm based on the natural gradient to design the controller of an open-loop stochastic distribution control system (SDCS) of multi-input and single output with a stochastic noise. Since the control input vector decides the shape of the output probability density function (PDF), the purpose of the controller design is to select a proper control input vector, so that the output PDF of the SDCS can be as close as possible to the target PDF. In virtue of the statistical characterizations of the SDCS, a new framework based on a statistical manifold is proposed to formulate the control design of the input and output SDCSs. Here, the Kullback–Leibler divergence is presented as a cost function to measure the distance between the output PDF and the target PDF. Therefore, an iterative descent algorithm is provided, and the convergence of the algorithm is discussed, followed by an illustrative example of the effectiveness.

Keywords: stochastic distribution control system; natural gradient algorithm; Kullback–Leibler divergence

1. Introduction

Information geometry [1–6] proposed by some scholars has been widely applied to various fields, such as neural network [7,8], control systems [9–12], dynamical system [13,14] and information science [15,16]. The main advantage of information geometry is that by considering the set of probability density functions (PDFs) as a manifold, one is able to investigate its properties geometrically. The parameters of a probability density function (PDF), regarded as the coordinate system of a statistical manifold, play important roles. As a classical example, consider the set S of normal PDFs with mean μ and variance σ^2, namely,

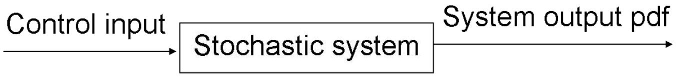
$$S = \left\{ p(x;\mu,\sigma) | p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \right\}.$$

Obviously, S can be viewed as a two-dimensional manifold, where (μ, σ) can be considered as a coordinate system. However, this statistical manifold is not a Euclidean space with respect to the Fisher metric, but a Riemannian manifold with a constant negative curvature. These observations in return enable us to investigate statistical properties from the viewpoint of information geometry.

There are many useful consequences of the information geometric theory; among them, the natural gradient algorithm [15] is well known. This algorithm has been applied to various stochastic models, such as stochastic network [16] and signal processing [17].

Stochastic distribution control systems (SDCSs) [18] were proposed from practical production systems, such as steel and paper making, and general material processing. The system is shown in Figure 1. The key point of the controller design problem is to formulate the control input, such that the output PDF is as close as possible to a required distribution shape.

Figure 1. The stochastic distribution control systems.



Generally, product quality data in industrial processes can be approximated by the Gaussian PDFs when the system operates normally. However, when abnormality occurs along the production line, these quality variables will not be Gaussian. Therefore, various iterative algorithms [19–22] have been presented to control the shape of PDFs for non-Gaussian cases. In [23], different kinds of SDCSs were discussed. One of them is the input and output model based output PDF control, which is investigated here. In this paper, we are mainly concerned with the information geometric algorithm to control the shape of PDFs. In [22], the authors firstly brought the idea of information geometry to the field of SDCSs and presented a comparative study on the parameter estimation performance between the geodesic equation and the B-spline function approximations, when the system output distributions are Gamma family distributions. Then, in [10] and [11], the authors generalized the results of [22] to more general cases where the system output distributions are assumed to be exponential family distributions and any regular distributions, respectively. There, the authors proposed information geometric algorithms using projection, natural gradient and geodesics, as well.

In the present paper, we investigate more complicated SDCSs of multi-input, single output with a stochastic noise from the viewpoint of information geometry. The remainder of this paper is organized as follows. In Section 2, we specify the SDCSs and re-describe them in the frame of information geometry. In Section 3, based on the natural gradient descent algorithm, a steepest descent algorithm is proposed from the viewpoint of information geometry. In Section 4, the convergence of the algorithm is discussed. In Section 5, an illustrative example is given.

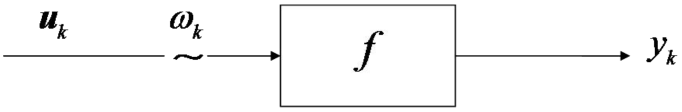
2. Model Description

In this paper, we investigate the open-loop SDCSs of multi-input and single output with a stochastic noise, where the structure of the systems is characterized by a known nonlinear function $f(\cdot)$ and the noise term is assumed to be subject to a known PDF $p_\omega(x)$. Therefore, the SDCSs can be expressed as:

$$y_k = f(\boldsymbol{u}_k, \omega_k), \tag{1}$$

where $\boldsymbol{u}_k = (u_k^1, \ldots, u_k^n) \in \mathbb{R}^n$ is the control input vector and $y_k \in \mathbb{R}^1$ is the output (see Figure 2).

Figure 2. The open-loop stochastic distribution control systems.



It is assumed that the function $f(\cdot)$ is invertible with respect to its noise term ω_k. Thus, according to $p_\omega(x)$ and Equation (1), the output PDFs of the system can be expressed by:

$$p(y;\boldsymbol{u}) = p_\omega\left(f^{-1}(y,\boldsymbol{u})\right)\frac{\partial f^{-1}(y,\boldsymbol{u})}{\partial y}. \tag{2}$$

This shows that Equation (2) implies how the control input vector $\boldsymbol{u}$ controls the shape of the output PDF of the SDCSs. For example, when the stochastic noise signal ω is subject to the normal distribution $N(0,1)$ and the stochastic distribution control system (SDCS) with single input and single output is formulated as $y = u^2 + \omega$, then the output PDF can be obtained as:

$$p(y;u) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(y-u^2)^2\right\}.$$

In order to guarantee the effectiveness, the following assumptions are required.

(1) The inverse function of $y = f(\boldsymbol{u},\omega)$ with respect to ω exists and is denoted by $\omega = f^{-1}(y,\boldsymbol{u})$, which is at least C^2 with respect to all variables $(y,\boldsymbol{u})$.

(2) The output PDF $p(y;\boldsymbol{u})$ is at least C^2 with respect to all variables $(y,\boldsymbol{u})$.

For the shape control of the PDF, the purpose of the controller design is to select the control input vector $\boldsymbol{u}_*$, so that $p(y;\boldsymbol{u}_*)$ is as close as possible to the target PDF $h(y)$. To formulate it in the frame of information geometry, we first define the relevant statistical manifold.

Definition 1. The statistical manifold S, called the system output manifold, is defined as:

$$S = \{p(y;\boldsymbol{u})\},$$

where $p(y;\boldsymbol{u})$ is in the form of Equation (2) and the control input vector $\boldsymbol{u} = (u^1,\ldots,u^n)^T \in \mathbb{R}^n$ plays the role of a coordinate system for S. Thus, S is an n-dimensional manifold.

Definition 2 ([5,6,24])**.** For a given statistical manifold, the Kullback–Leibler divergence between two points P and Q corresponding to the PDF $p(x)$ and the PDF $q(x)$, respectively, is defined by:

$$J(P,Q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x.$$

Notice that the Kullback–Leibler divergence neither satisfies the triangular inequality nor is symmetric. Hence, it is not a usual distance function. However, the Kullback–Leibler divergence between two neighboring points θ and $\theta + \mathrm{d}\theta$ can be approximated by using the Fisher metric:

$$J(\theta, \theta + \mathrm{d}\theta) \approx J(\theta + \mathrm{d}\theta, \theta) \approx \frac{1}{2} g_{ij}(\theta) \, \mathrm{d}\theta^i \, \mathrm{d}\theta^j,$$

where the terms of order $O(|\,\mathrm{d}\theta|^3)$ are neglected. Thus, the Kullback–Leibler divergence is a distance-like measure of two points on a statistical manifold and has been widely applied, for example, to information theory. Here, (g_{ij}) is the Fisher metric equipped on manifold S, whose components are expressed as:

$$g_{ij} = E[\partial_i l(x;\theta) \;\; \partial_j l(x;\theta)], \;\; i,j = 1,2,\ldots,n, \tag{3}$$

where $l(x;\theta) = \log p(x;\theta)$, $\partial_i = \frac{\partial}{\partial \theta^i}$ and E denotes the expectation with respect to $p(x;\theta)$. Manifold (S,g) is hence a Riemannian manifold.

Here, we use the following Kullback–Leibler divergence function to measure the difference between $h(y)$ and $p(y;\boldsymbol{u})$ by:

$$J(\boldsymbol{u}) = \int h(y) \log \frac{h(y)}{p(y;\boldsymbol{u})} \, \mathrm{d}y, \tag{4}$$

that is, $J(\boldsymbol{u})$ is considered a cost function. Our purpose is to design a controller so that the control input vector $\boldsymbol{u}_*$ minimizes $J(\boldsymbol{u})$, namely,

$$\boldsymbol{u}_* = \arg\min_{\boldsymbol{u}_k} J(\boldsymbol{u}), \qquad k = 1,2,\ldots.$$

Alternatively, the problem can be re-described as selecting the points $p(y;\boldsymbol{u})$ on S with the coordinate system $\boldsymbol{u}$ to make the points as close as possible to the target point $h(y)$ in the frame of information geometry.

3. Natural Gradient Algorithm

In this section, we will introduce an iterative algorithm for the controller design from the viewpoint of information geometry. It is a fact that the ordinary gradient method is a popular learning method in Euclidean space. However, most practical problems are non-Euclidean, where the ordinary gradient method loses its effectiveness. In such cases, the ordinary gradient does not give the steepest descent direction of a target function, but the natural gradient does. Next, we will introduce an important lemma about the natural gradient.

Let $\{\omega \in \mathbb{R}^n\}$ be a parameter space on which a function L is defined.

Lemma 1 ([15])**.** *The steepest descent direction of $L(\omega)$ on a Riemannian manifold is given by:*

$$-\tilde{\nabla} L(\omega) = -G^{-1}(\omega) \nabla L(\omega), \tag{5}$$

where $G^{-1} = (g^{ij})$ is the inverse of the Riemannian metric $G = (g_{ij})$ and $\nabla L(\omega)$ is the ordinary gradient:

$$\nabla L(\omega) = \left(\frac{\partial}{\partial \omega_1} L(\omega), \frac{\partial}{\partial \omega_2} L(\omega), \ldots, \frac{\partial}{\partial \omega_{n-1}} L(\omega) \right).$$

To obtain the steepest descent algorithm, we first formulate the Fisher metric of S as follows.

Proposition 1. *The components of the Fisher metric of S are given by:*

$$\begin{aligned} g_{ij} = & \int \frac{\frac{\partial f^{-1}(y,\mathbf{u})}{\partial y}}{p_\omega(f^{-1}(y,\mathbf{u}))} \frac{\partial p_\omega(f^{-1}(y,\mathbf{u}))}{\partial u^i} \frac{\partial p_\omega(f^{-1}(y,\mathbf{u}))}{\partial u^j} \mathrm{d}y \\ & + \int \left(\frac{\partial p_\omega(f^{-1}(y,\mathbf{u}))}{\partial u^i} \frac{\partial^2 f^{-1}(y,\mathbf{u})}{\partial y \partial u^j} + \frac{\partial p_\omega(f^{-1}(y,\mathbf{u}))}{\partial u^j} \frac{\partial^2 f^{-1}(y,\mathbf{u})}{\partial y \partial u^i} \right) \mathrm{d}y \\ & + \int \frac{p_\omega(f^{-1}(y,\mathbf{u}))}{\frac{\partial f^{-1}(y,\mathbf{u})}{\partial y}} \frac{\partial^2 f^{-1}(y,\mathbf{u})}{\partial y \partial u^i} \frac{\partial^2 f^{-1}(y,\mathbf{u})}{\partial y \partial u^j} \mathrm{d}y, \end{aligned}$$

for $i, j \in \{1, 2, \ldots, n\}$.

Proof. The first order derivatives of $\log p(y; \mathbf{u})$ with respect to u^i $(i = 1, 2, \ldots, n)$ are given by:

$$\frac{\partial \log p(y;\mathbf{u})}{\partial u^i} = \frac{1}{p_\omega(f^{-1}(y,\mathbf{u}))} \frac{\partial p_\omega(f^{-1}(y,\mathbf{u}))}{\partial u^i} + \frac{1}{\frac{\partial f^{-1}(y,\mathbf{u})}{\partial y}} \frac{\partial^2 f^{-1}(y,\mathbf{u})}{\partial y \partial u^i}. \tag{6}$$

Note that $\frac{\partial \log p(y;\mathbf{u})}{\partial u^i}$ must satisfy the condition:

$$E\left[\frac{\partial \log p(y;\mathbf{u})}{\partial u^i} \right] = 0,$$

for all $i = 1, 2, \ldots, n$. Combining (3) and (6), we obtain the conclusion in Proposition 1. This completes the proof. □

Thus, we have the following iterative descent algorithm.

Theorem 1. *Based on the natural gradient algorithm, the steepest descent algorithm for the control input vector $\mathbf{u}$ of the considered stochastic distribution control systems is given by:*

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \frac{\varepsilon^2}{2\lambda} G_k^{-1} \nabla J(\mathbf{u}_k), \tag{7}$$

where G_k^{-1} is the inverse of the Fisher metric $G_k = G|_{\mathbf{u}=\mathbf{u}_k}$, and ε is a sufficiently small positive constant, which determines the step size. Here, we set:

$$\nabla J(\mathbf{u}_k) = \left(\frac{\partial J(\mathbf{u})}{\partial u^1}, \ldots, \frac{\partial J(\mathbf{u})}{\partial u^n} \right)^T \Big|_{u=u_k}$$

and:

$$\lambda = \frac{\varepsilon}{2} \sqrt{\nabla J(\mathbf{u}_k)^T G_k^{-1} J(\mathbf{u}_k)}.$$

Proof. Let P_k and P_{k+1} be two close points on S corresponding to the functions $\log p(y; \mathbf{u}_k)$ and $\log p(y; \mathbf{u}_{k+1})$, whose coordinates are given by $\mathbf{u}_k = (u_k^1, \ldots, u_k^n)^T$ and $\mathbf{u}_{k+1} = \mathbf{u}_k + \triangle \mathbf{u}_k$, respectively, where $\mathbf{u}_{k+1} = (u_{k+1}^1, \ldots, u_{k+1}^n)^T$, and $\triangle \mathbf{u}_k = (\triangle u_k^1, \ldots, \triangle u_k^n)^T$. Therefore, our purpose is to formulate an iterative formula with respect to $\mathbf{u}_{k+1}$. Assume that the vector $\overrightarrow{P_k P_{k+1}} \in T_{P_k} S$ has a fixed length, namely,

$$|\overrightarrow{P_k P_{k+1}}|^2 = \varepsilon^2,$$

where ε is a sufficiently small positive constant. Then, we put:

$$\overrightarrow{P_k P_{k+1}} = \varepsilon \mathbf{v}, \tag{8}$$

where:

$$\mathbf{v} = a^i \frac{\partial \log p(y; \mathbf{u})}{\partial u^i} \in T_{P_k} S \tag{9}$$

can be considered as a tangent vector of $T_{P_k} S$ at P_k. We denote $\mathbf{a} = (a^1, \ldots, a^n)^T$, and the tangent vector $\mathbf{v}$ satisfies:

$$|\mathbf{v}|^2 = \langle \mathbf{v}, \mathbf{v} \rangle = \mathbf{a}^T G_k \mathbf{a} = 1, \tag{10}$$

where G_k means that $G|_{u=u_k}$.

To reveal the iterative relation of the functions $\log p(y; \mathbf{u}_k)$ and $\log p(y; \mathbf{u}_{k+1})$ between the sample times k and $k+1$, the following equation is performed approximately:

$$\log p(y; \mathbf{u}_{k+1}) - \log p(y; \mathbf{u}_k) = (\mathbf{u}_{k+1} - \mathbf{u}_k)^T \nabla \log p(y; \mathbf{u}_k), \tag{11}$$

where:

$$\nabla \log p(y, \mathbf{u}_k) = \left(\frac{\partial \log p(y; \mathbf{u})}{\partial u^1}, \ldots, \frac{\partial \log p(y; \mathbf{u})}{\partial u^n} \right)^T \Big|_{u=u_k}.$$

Combining Equations (8), (9) and (11), we get the following equation:

$$\triangle \mathbf{u}_k = \varepsilon \mathbf{a}. \tag{12}$$

From Equation (12), we get the relations between $J(\mathbf{u}_{k+1})$ and $J(\mathbf{u}_k)$ as follows:

$$\begin{aligned} J(\mathbf{u}_{k+1}) =& J(\mathbf{u}_k) + (\mathbf{u}_{k+1} - \mathbf{u}_k)^T \nabla J(\mathbf{u}_k) \\ =& J(\mathbf{u}_k) + \varepsilon \mathbf{a}^T \nabla J(\mathbf{u}_k), \end{aligned} \tag{13}$$

where:

$$\nabla J(\mathbf{u}_k) = \left(\frac{\partial J(\mathbf{u})}{\partial u^1}, \ldots, \frac{\partial J(\mathbf{u})}{\partial u^n} \right)^T \Big|_{u=u_k}.$$

Note that $\mathbf{u}_k$ is known at the sample time $k+1$. Here, $\mathbf{a} = (a^1, \ldots, a^n)^T$ should be selected, such that the following performance function:

$$F(a^1, \ldots, a^n) = J(\mathbf{u}_k) + \varepsilon \mathbf{a}^T \nabla J(\mathbf{u}_k) + \lambda \mathbf{a}^T G_k \mathbf{a} \tag{14}$$

is minimized, where the first two terms are the linear approximation of the Kullback–Leibler divergence $J(\boldsymbol{u}_k)$ at the sample time k, while the third term is a natural quadratic constraint for $\mathbf{a} = (a^1, \ldots, a^n)^T$. Then, the optimal vector **a** can be obtained as:

$$\mathbf{a} = -\frac{\varepsilon}{2\lambda} G_k^{-1} \nabla J(\boldsymbol{u}_k). \tag{15}$$

From above, it can be seen that Equation (15) sets up the necessary condition for an optimal vector **a**. Now, let us consider the sufficient condition of Equation (15) to minimize the performance function (14).

Firstly, we determine the value of the parameter λ. Since:

$$\mathbf{a}^T G \mathbf{a} = \frac{\varepsilon^2}{4\lambda^2} \nabla J(\boldsymbol{u}_k)^T G_k^{-1} G_k G_k^{-1} \nabla J(\boldsymbol{u}_k) = \frac{\varepsilon^2}{4\lambda^2} \nabla J(\boldsymbol{u}_k)^T G_k^{-1} \nabla J(\boldsymbol{u}_k) = 1,$$

and $\nabla J(\boldsymbol{u}_k)^T G_k^{-1} \nabla J(\boldsymbol{u}_k)$ is positive, we get the value of λ as:

$$\lambda = \frac{\varepsilon}{2} \sqrt{J(\boldsymbol{u}_k)^T G_k^{-1} \nabla J(\boldsymbol{u}_k)}.$$

Then, the Hessian matrix of $F(a^1, \ldots, a^n)$ with respect to the vector $\mathbf{a} = (a^1, \ldots, a^n)^T$ is given by:

$$H = \begin{pmatrix} \frac{\partial^2 F}{\partial a^1 \partial a^1} & \frac{\partial^2 F}{\partial a^1 \partial a^2} & \cdots & \frac{\partial^2 F}{\partial a^1 \partial a^n} \\ \frac{\partial^2 F}{\partial a^2 \partial a^1} & \frac{\partial^2 F}{\partial a^2 \partial a^2} & \cdots & \frac{\partial^2 F}{\partial a^2 \partial a^n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 F}{\partial a^n \partial a^1} & \frac{\partial^2 F}{\partial a^n \partial a^2} & \cdots & \frac{\partial^2 F}{\partial a^n \partial a^n} \end{pmatrix} = 2\lambda G_k.$$

since λ is positive and G_k is positive definite, the Hessian matrix is positive definite. This guarantees that the vector **a** in the form of Equation (15) minimizes the performance function (14) naturally.

Substituting Equation (15) into Equation (12), we obtain the conclusion in Equation (7), which gives a steepest descent direction algorithm for this stochastic distribution control problem. □

We summarize the algorithm above as follows:

(1) Initialize u_0.

(2) At the sample time $k-1$, formulate $\nabla J(\boldsymbol{u}_{k-1})$ and use Equation (1) to give the inverse G_{k-1}^{-1} of the Fisher metric G_{k-1}.

(3) Calculate $\boldsymbol{u}_k$ using Equation (7) and apply it to the stochastic system.

(4) If $J(\boldsymbol{u}_k) < \delta$, where δ is a positive constant, which is determined by the precision needed, escape. Additionally, at the sample time k, the output PDF $p(y; \boldsymbol{u}_k)$ is the final one. If not, turn to Step 5.

(5) Increase k by one and go back to Step 2.

4. Convergence of the Algorithm

Now, let us study the convergence of the algorithm for the control input vector $\boldsymbol{u}$ in Equation (7).

Lemma 2. *For an n-dimensional manifold M with a distance function $d(x, y)$, the norm is defined by $\|x - y\| = d(x, y)$. Let $f : M \to \mathbb{R}^n$ be a continuous mapping on a compact set D of M and the set $\Omega = \{x \in D |\ f(x) = 0\}$ be finite. If the sequence $\{x^m\}_{m=1}^{\infty} \subset D$ satisfies:*

$$\lim_{m\to\infty} \|x^{m+1} - x^m\| = 0 \text{ and } \lim_{m\to\infty} \|f(x^m)\| = 0,$$

then there exists an $x^* \in \Omega$*, such that:*

$$\lim_{m\to\infty} x^m = x^*.$$

Proof. Let:

$$B(x,\epsilon) = \{y \in M| \, \|x-y\| < \epsilon, \epsilon > 0\}$$

and:

$$\Omega = \{x \in D| \, f(x) = 0\} = \{a^1, a^2, ..., a^s| \, s \in \mathbb{N}_+\}.$$

First, we shall prove the following conclusion: for any $\epsilon > 0$, there exists a constant $K > 0$, such that $x^m \in \bigcup\limits_{i=1}^{s} B(a^i, \epsilon)$ for arbitrary $m > K$.

Now, we give the proof by contradiction. If for a certain $\epsilon_0 > 0$, we have that for arbitrary $K > 0$, there exists an $m > K$, such that $x^m \notin \bigcup\limits_{i=1}^{s} B(a^i, \epsilon_0)$, then, for $K = 1$, we get an $m_1 > 1$ satisfying $x^{m_1} \notin \bigcup\limits_{i=1}^{s} B(a^i, \epsilon_0)$. Moreover, for $K = m_1$, we get an $m_2 > m_1$, such that $x^{m_2} \notin \bigcup\limits_{i=1}^{s} B(a^i, \epsilon_0)$. Following this way, we get a subsequence $\{x^{m_j}\}$ of $\{x^m\}$, satisfying $x^{m_j} \notin \bigcup\limits_{i=1}^{s} B(a^i, \epsilon_0)$ for arbitrary j.

Since D is compact, $\{x^{m_j}\}$ must have a convergent subsequence $\{x^{m_{j_i}}\}$, namely,

$$\lim_{i\to\infty} x^{m_{j_i}} = \bar{x} \in \overline{D - \bigcup_{i=1}^{s} B(a^i, \epsilon_0)}.$$

It is obvious that:

$$\bar{x} \notin \bigcup_{i=1}^{s} B(a^i, \frac{\epsilon_0}{2}) \text{ and } \|f(\bar{x})\| \neq 0.$$

As f is continuous, we have:

$$\lim_{i\to\infty} \|f(x^{m_{j_i}})\| = \|f(\bar{x})\| \neq 0,$$

which contradicts $\lim\limits_{m\to\infty} \|f(x^m)\| = 0$. Therefore, the conclusion in the beginning holds.

Since $\{x^m\}_{m=1}^{\infty} \subset D$, we can get a convergent subsequence $\{x^{m_k}\}_{k=1}^{\infty}$. Setting $\lim\limits_{k\to\infty} x^{m_k} = x^*$, it is trivial that $x^* \in \Omega$ according to the process of the conclusion above.

Because the set Ω is finite, setting $\epsilon_0 = \frac{1}{2} \min\limits_{1\leq i,j\leq s} \{\|a^i - a^j\| \mid i \neq j\}$, we get $B(a^i, \epsilon_0) \cap B(a^j, \epsilon_0) = \emptyset$, when $i \neq j$.

Moreover, we shall prove $\lim\limits_{m\to\infty} x^m = x^*$ by its equivalent proposition: for arbitrary $0 < \epsilon < \frac{\epsilon_0}{2}$, there exists a $K > 0$, such that $x^m \in B(x^*, \epsilon)$ for any $m \geq K$.

For arbitrary $0 < \epsilon < \frac{\epsilon_0}{2}$, we get that there exists a $K_1 > 0$, such that:

$$x^m \in \bigcup_{i=1}^{s} B(a^i, \epsilon), \tag{16}$$

for arbitrary $m > K_1$. Meanwhile, we also have:

$$\|\alpha - \beta\| > \epsilon_0, \tag{17}$$

where $\alpha \in B(a^i, \epsilon)$ and $\beta \in B(a^j, \epsilon)$ are arbitrary, when $i \neq j$.

As $\lim_{k\to\infty} x^{m_k} = x^*$, there exists a constant $L > 0$, so that:

$$x^{m_k} \in B(x^*, \epsilon), \tag{18}$$

when $k > L$.

Since $\lim_{m\to\infty} \|x^{m+1} - x^m\| = 0$, for $\epsilon_0 > 0$, there exists a $K_2 > 0$, such that:

$$\|x^{m+1} - x^m\| < \epsilon_0, \tag{19}$$

for arbitrary $m > K_2$.

Take $\overline{K} = \max\{K_1, K_2, m_L\}$; then, we set $K = \min_k\{m_k \mid m_k \geq \overline{K}\}$, so that $x^K \in \{x^{m_k}\}$.

Finally, we shall finish the proof by using mathematical induction.

When $m = K$, since $x^K \in \{x^{m_k}\}$, we get $x^m \in B(x^*, \epsilon)$ from Equation (18) directly.

If when $m = N \geq K$, $x^N \in B(x^*, \epsilon)$, then when $m = N + 1$, we see that x^{N+1} should be contained in the union of the s open balls from Equation (16), while from Equations (17) and (19), we also get that x^N and x^{N+1} must be in the same ball, *i.e.*, $x^{N+1} \in B(x^*, \epsilon)$.

This finishes the proof of Lemma 2. □

Lemma 3. *Let $J(\mathrm{u})$ be at least C^2 with respect to u. For an initial value $\mathrm{u}_0 \in \mathbb{R}^n$, suppose that the level set $L = \{\mathrm{u} \in \mathbb{R}^n | J(\mathrm{u}) \leqslant J(\mathrm{u}_0)\}$ is compact. The sequence $\{\mathrm{u}_k\}$ in Equation (7) has the following property: for a certain k_0, either $G_{k_0}^{-1}\nabla J(\mathrm{u}_{k_0}) = 0$ or when $k \to \infty$, $G_k^{-1}\nabla J(\mathrm{u}_k) \to 0$, where G_k^{-1} is the inverse of the Fisher metric G_k.*

Proof. Set $c_k = G_k^{-1}\nabla J(\mathrm{u}_k)$ and $\alpha = \frac{\varepsilon^2}{2\lambda}$ for simplicity.

Assume that $c_k \neq 0$ for any sample time k. Now, let us give a proof by contradiction. Suppose that when $k \to \infty$, $c_k \to 0$ does not hold, that is, there exists an $\varepsilon_0 > 0$, so that the norm of c_k satisfies:

$$\|c_k\| \geqslant \varepsilon_0 \tag{20}$$

for infinitely many k, where $\|c_k\|^2 = c_k^T G_k c_k$. Thus, for such k, Equation (20) can be rewritten as:

$$\frac{c_k^T G_k c_k}{\|c_k\|} \geqslant \varepsilon_0. \tag{21}$$

Then, from the mean value theorem of differentials, we get:

$$\begin{aligned}
J(\mathrm{u}_k - \alpha c_k) =& J(\mathrm{u}_k) - \alpha c_k^T \nabla J(\mathrm{v}_k) \\
=& J(\mathrm{u}_k) - \alpha c_k^T \nabla J(\mathrm{u}_k) - \alpha c_k^T \Big(\nabla J(\mathrm{v}_k) - \nabla J(\mathrm{u}_k)\Big) \\
=& J(\mathrm{u}_k) - \alpha c_k^T G_k G_k^{-1} \nabla J(\mathrm{u}_k) - \alpha c_k^T G_k G_k^{-1} \Big(\nabla J(\mathrm{v}_k) - \nabla J(\mathrm{u}_k)\Big) \\
=& J(\mathrm{u}_k) - \alpha c_k^T G_k c_k - \alpha c_k^T G_k \Big(c(\mathrm{v}_k) - c_k\Big) \\
=& J(\mathrm{u}_k) + \alpha \|c_k\| \frac{(-c_k^T) G_k c_k}{\|c_k\|} + \alpha(-c_k^T) G_k \Big(c(\mathrm{v}_k) - c_k\Big) \\
\leqslant& J(\mathrm{u}_k) + \alpha \|c_k\| \frac{(-c_k^T) G_k c_k}{\|c_k\|} + \alpha \|c_k\| \|c(\mathrm{v}_k) - c_k\| \\
=& J(\mathrm{u}_k) + \alpha \|c_k\| \left(\frac{(-c_k^T) G_k c_k}{\|c_k\|} + \|c(\mathrm{v}_k) - c_k\|\right),
\end{aligned} \tag{22}$$

where $c(u) = G^{-1}(u)\nabla J(u)$, and v_k belongs to the continuous space between u_k and $u_k - \alpha c_k$.

Since $c(u)$ is continuous and the level set L is compact, $c(u)$ is uniformly continuous on L, which means that there exists a $\beta > 0$, when $0 \leqslant \|u_k - \alpha c_k - u_k\| = \|\alpha c_k\| \leqslant \beta$,

$$\|c(v_k) - c_k\| \leqslant \frac{1}{2}\varepsilon_0 \tag{23}$$

holds for all the k.

Then, taking $\alpha = \frac{\beta}{\|c_k\|}$ in Equation (22) and combining Equation (21) with Equation (23), we have:

$$\begin{aligned} J(u_{k+1}) =& J(u_k - \alpha c_k) = J(u_k - \frac{\beta}{\|c_k\|}c_k) \\ \leqslant& J(u_k) + \beta\left(\frac{(-c_k^T)G_k c_k}{\|c_k\|} + \|c(v_k) - c_k\|\right) \\ \leqslant& J(u_k) + \beta(-\varepsilon_0 + \frac{1}{2}\varepsilon_0) \\ =& J(u_k) - \frac{1}{2}\beta\varepsilon_0 \end{aligned}$$

for infinitely many k.

On the other hand, since:

$$J(u_k) - J(u_{k-1}) = -\frac{\varepsilon^2}{2\lambda}\nabla J(u_{k-1})^T G_{k-1}^{-1}\nabla J(u_{k-1}),$$

in which G_{k-1}^{-1} is positive definite and $\lambda > 0$, we have $J(u_k) - J(u_{k-1}) < 0$, *i.e.*, $\{J(u_k)\}$ is monotone decreasing with respect to k.

The level set L is compact, which implies that $\lim_{k\to\infty} J(u_k)$ exists, namely,

$$J(u_k) - J(u_{k-1}) \to 0,$$

when $k \to \infty$.

This is a contradiction. This completes the proof of Lemma 3. □

Theorem 2. *Let $\nabla J(u)$ be a continuous function with the compact level set L, and suppose the set $\Omega = \{u \in L|\nabla J(u) = 0\}$ is finite. Then, there exists $u_* \in \Omega$, such that:*

$$\lim_{k\to\infty} u_k = u_*.$$

Proof. From Lemma 3, we get:

$$\lim_{k\to\infty} \|u_{k+1} - u_k\| = 0. \tag{24}$$

Meanwhile, similarly with the process of the proof of Lemma 3, we have:

$$\lim_{k\to\infty} \|\nabla J(u_k)\| = 0. \tag{25}$$

Therefore, from Equations (24), (25) and Lemma 2, we get the conclusion in Theorem 2. □

5. Simulations

The dynamic characteristic of a simple nonlinear stochastic system is considered as:

$$y_{k+1} = (\omega_k \mu_k - \sigma_k)^{\frac{1}{3}},$$

where $\omega_k \in [0, +\infty)$ and (μ, σ) is the input vector. Here, the stochastic noise ω_k is a random process whose PDF is written as:

$$p_\omega(x) = \frac{1}{3750} x^3 e^{-\frac{1}{5}x},$$

where $x \in [0, +\infty)$.

The target PDF $h(y)$ is given by:

$$h(y) = \begin{cases} -\frac{2}{33}(y^2 - 5y) & y \in [1, 4], \\ 0 & \text{else.} \end{cases}$$

Then, from Equation (2), we can get the output PDF $p(y; \mu, \sigma)$ as:

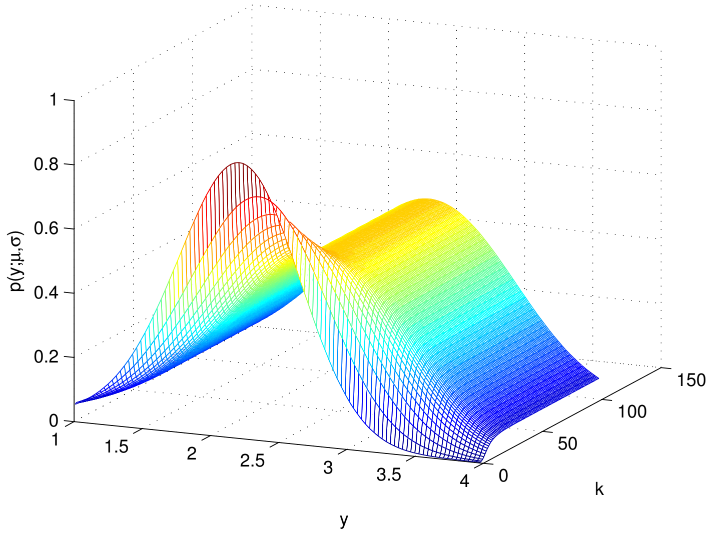
$$p(y; \mu, \sigma) = \frac{y^2\,(y^3 + \sigma)^3}{1250\mu^4} e^{-\frac{1}{5\mu}(y^3 + \sigma)}.$$

The components of the Fisher metric can hence be obtained as:

$$\begin{aligned} g_{11} =& \frac{1}{3\mu^2} e^{-\frac{64+\sigma}{5\mu}} \left(\frac{(64+\sigma)^4}{625\mu^4} + \frac{6(64+\sigma)^3}{125\mu^3} + \frac{6(64+\sigma)^2}{25\mu^2} + \frac{12(64+\sigma)}{5\mu} + 12 \right) \\ & - \frac{1}{3\mu^2} e^{-\frac{1+\sigma}{5\mu}} \left(\frac{(1+\sigma)^4}{625\mu^4} + \frac{6(1+\sigma)^3}{125\mu^3} + \frac{6(1+\sigma)^2}{25\mu^2} + \frac{12(1+\sigma)}{5\mu} + 12 \right), \\ g_{12} =& g_{21} = \frac{125}{3\mu^2} e^{-\frac{64+\sigma}{5\mu}} \left(-\frac{(64+\sigma)^3}{125\mu^3} + \frac{3(64+\sigma)^2}{25\mu^2} - \frac{6(64+\sigma)}{5\mu} - 6 \right) \\ & - \frac{125}{3\mu^2} e^{-\frac{1+\sigma}{5\mu}} \left(-\frac{(1+\sigma)^3}{125\mu^3} + \frac{3(1+\sigma)^2}{25\mu^2} - \frac{6(1+\sigma)}{5\mu} - 6 \right), \\ g_{22} =& - \frac{25\sigma}{\mu^2} e^{-\frac{64+\sigma}{5\mu}} \left(\frac{2(64+\sigma)^2}{5\mu} + \frac{(20\mu - \sigma)(64+\sigma)}{5\mu} + 20\mu - \sigma \right) \\ & + \frac{25\sigma}{\mu^2} e^{-\frac{1+\sigma}{5\mu}} \left(\frac{2(1+\sigma)^2}{5\mu} + \frac{(20\mu - \sigma)(1+\sigma)}{5\mu} + 20\mu - \sigma \right). \end{aligned}$$

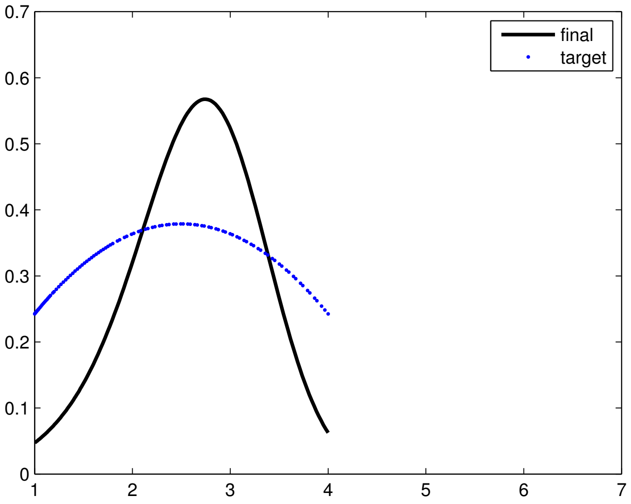
To start the simulation, the initial value is chosen as $u_0 = (\mu_0, \sigma_0)^T = (0.7, 2.5)^T$. The weights ε and λ are taken as 0.6 and 0.8, respectively. As a result, the response of the output PDFs is shown in Figure 3, in which y denotes the output of the system, $p(y; \mu, \sigma)$ denotes the PDF of the output y and k denotes the sample time.

Figure 3. The response of the output probability density functions.



In order to illustrate the effectiveness of the control law in detail, the comparison between the final controlled output PDF and the target function is shown in Figure 4, where the horizontal axis denotes the output y and the vertical axis denotes the PDF $p(y; \mu, \sigma)$ of y.

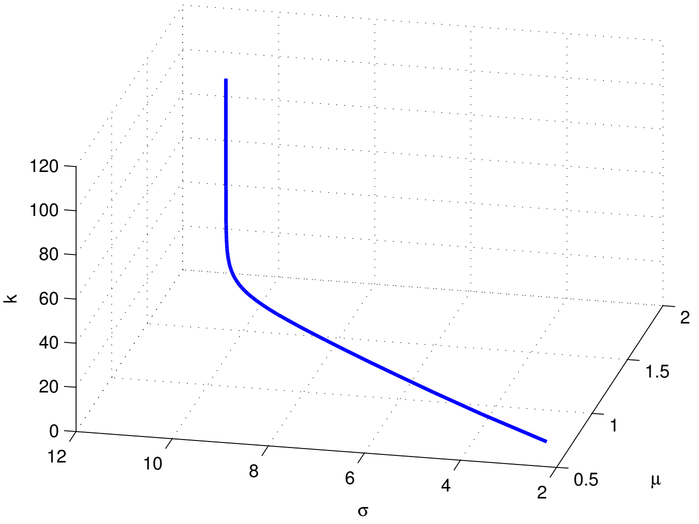
Figure 4. The final and the target probability density functions.



Obviously, the output PDF can be controlled to be a steady state shown in Figure 4. However, as we know, assume that A and B are two sets in a metric space (e.g., Euclidean space) with the distance function d; then, the distance $d(A, B) = \max_{x\in A, y\in B} d(x, y)$ may be larger than zero. Actually, in our simulation, the target PDF is in the set of second order polynomials, and the PDF $p(y; \mu, \sigma)$ of the output y is exponential. Therefore, the non-zero steady error still exists all of the time.

The response of the optimal control input sequences is shown in Figure 5, in which (μ, σ) denotes the input vector and k denotes the sample time.

Figure 5. The optimal control input sequences.



From above, it can be concluded that the simulation results demonstrate the effectiveness of the presented method, which gives a solution to control the shape of the output PDF.

6. Conclusions

In this paper, we investigate the open-loop stochastic distribution control systems of multi-input and single output with a stochastic noise, via the advantage of information geometric theory.

(1) By the statistical characterizations of the stochastic distribution control systems, we formulate the controller design in the frame of information geometry. By virtue of the natural gradient algorithm, a steepest descent algorithm is proposed.
(2) The convergence of the obtained algorithm is proven.
(3) An example is discussed in detail to demonstrate our algorithm.

Acknowledgments

The authors would like to express their sincere thanks to the referees for their valuable suggestions.

This subject are supported by the National Natural Science Foundations of China, Nos. 10871218, 10932002, 61179031, and the Doctoral Program of Higher Education of China through Grant, No. 20131103120027. The work of Linyu Peng was supported by the Math Department of the University of Surrey and the China Scholarship Council. The work of Linyu Peng is also supported by JST-CREST.

Author Contributions

In this paper, Zhenning Zhang is in charge of the control theory and information geometric theory, Huafei Sun is in charge of the geometric theory and paper writing, Linyu Peng is in charge of the geometric theory and the simulation, and Lin Jiu is in charge of the topological theory. The authors have read and approved the final published manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Rao, C.R. Infromation and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta. Math. Soc.* **1945**, *37*, 81–91.
2. Efron, B. Defining the curvature of a statistical problem. *Ann. Stat.* **1975**, *3*, 1189–1242.
3. Efron, B. The geometry of exponential families. *Ann. Stat.* **1978**, *6*, 362–376.
4. Chentsov, N.N. *Statistical Decision Rules and Optimal Inference*; AMS: Providence, RI, USA, 1982.
5. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Oxford University Press: Oxford, UK, 2000.
6. Amari, S. *Differential Geometrical Methods in Statistics*; Springer-Verlag: Berlin/Heidelberg, Germany, 1990.
7. Amari, S. Information geometry of the EM and em algorithm for neural networks. *Neural Netw.* **1995**, *8*, 1379–1408.
8. Amari, S.; Kurata, K.; Nagaoka, H. Information geometry of Boltzmann machines. *IEEE Trans. Neural Netw.* **1992**, *3*, 260–271.
9. Amari, S. Differential geometry of a parametric family of invertible linear systems-Riemannian metric, dual affine connections, and divergence. *Math. Syst. Theory* **1987**, *20*, 53–83.
10. Zhang, Z.; Sun, H.; Zhong, F. Natural gradient-projection algorithm for distribution control. *Optim. Control Appl. Methods* **2009**, *30*, 495–504.
11. Zhong, F.; Sun, H.; Zhang, Z. An Information geometry algorithm for distribution control. *Bull. Braz. Math. Soc.* **2008**, *39*, 1–10.
12. Zhang, Z.; Sun, H.; Peng, L. Natural gradient algorithm for stochastic distribution systems with output feedback. *Differ. Geom. Appl.* **2013**, *31*, 682–690.
13. Peng, L.; Sun, H.; Sun, D.; Yi, J. The geometric structures and instability of entropic dynamical models. *Adv. Math.* **2011**, *227*, 459–471.
14. Peng, L.; Sun, H.; Xu, G. Information geometric characterization of the complexity of fractional Brownian motions. *J. Math. Phys.* **2012**, *53*, 123305.
15. Amari, S. Natural gradient works efficiently in learning. *Neural Comput.* **1998**, *10*, 251–276.
16. Amari, S. Natural gradient learning for over- and under-complete bases in ICA. *Neural Comput.* **1999**, *11*, 1875–1883.
17. Park, H.; Amari, S.; Fukumizu, K. Adaptive natural gradient learning algorithms for various stochastic model. *Neural Netw.* **2000**, *13*, 755–764.
18. Guo, L.; Wang, H. *Stochastic Distribution Control System Design: A Convex Optimization Approach*; Springer: London, UK, 2010.
19. Wang, H. Control of Conditional output probability density functions for general nonlinear and non-Gaussian dynamic stochastic systems. *IEE Proc. Control Theory Appl.* **2003**, *150*, 55–60.

20. Guo, L.; Wang, H. Minimum entropy filtering for multivariate stochastic systems with non-Gaussian noises. *IEEE Trans. Autom. Control* **2006**, *51*, 695–670.
21. Wang, A.; Afshar, P.; Wang, H. Complex stochastic systems modelling and control via iterative machine learning. *Neurocomputing* **2008**, *71*, 2685–2692.
22. Dodson, C.T.J.; Wang, H. Iterative approximation of statistical distributions and relation to information geometry. *Stat. Inference Stoch. Process.* **2001**, *4*, 307–318.
23. Wang, A.; Wang, H.; Guo, L. Recent Advances on Stochastic Distribution Control: Probability Density Function Control. In Proceedings of the CCDC 2009: Chinese Control and Decision Conference, Guilin, China, 17–19 June 2009; doi: 10.1109/CCDC.2009.5195154.
24. Sun, H.; Peng, L.; Zhang, Z. Information geometry and its applications. *Adv. Math. (China)* **2011**, *40*, 257–269. (In Chinese)