Reparametrization-covariant theory for on-line learning of probability distributions

Toshiaki Aida
Tokyo Metropolitan College of Aeronautical Engineering, Minami-senjyu, Arakawa-ku, Tokyo 116-0003, Japan
(Received 30 April 2001; published 30 October 2001)

We discuss the on-line learning of probability distributions in a reparametrization covariant formulation. Reparametrization covariance plays an essential role not only to respect an intrinsic property of ''information'' but also for pattern recognition problems. We can obtain an optimal on-line learning algorithm with reparametrization invariance, where the conformal gauge connects a covariant formulation with a noncovariant one in a natural way.

DOI: 10.1103/PhysRevE.64.056128 PACS number(s): 02.50.Fz, 11.10.Hi, 84.35.+i, 89.70.+c

I. INTRODUCTION

The inference of probability distributions is of fundamental importance in learning theory. For example, it gives us a unified framework including both supervised and nonsupervised learning schemes. One approach to the problem is to project the distributions to a space parametrized by a finite set of coordinates and to determine their locations on it. However, this prescription inevitably excludes the distributions with a lot of features. Therefore, efforts have been devoted to the original problem of determining probability distributions themselves.

Among others, field theory formulates the problem in a natural way. It is the scaling analysis that leads us to optimal algorithms not only in the batch learning scheme but also in the on-line learning one [1,2]. This is because the scaling tells us how to control the number of degrees of freedom that we prepare for a learning system. This problem has been approached from various directions [3].

However, the algorithms of Refs. [1,2] depend on a specific coordinate, in which we observe example data. Information or probability distributions are originally independent of how to describe them. Therefore, it is desirable to derive learning algorithms that respect this intrinsic property of information. We can resolve the problem in a geometrically covariant formulation, which has already been obtained in the context of batch learning [4]. It is reparametrization invariance that characterizes such algorithms. Reparametrization invariance is also important from the practical point of view. It plays a crucial role in pattern recognition problems such as visual and speech information processes.

In this paper, we discuss on-line learning of probability distributions in a reparametrization-covariant form. We will find that the conformal gauge plays a crucial role in connecting a covariant formulation with a noncovariant one. Therefore, we can also obtain an optimal on-line learning algorithm with reparametrization invariance.

II. COVARIANT FORMULATION

Let us describe inference problems in such a way as is independent from a specific coordinate system. We observe a series of sample points $P_1, P_2, \ldots, P_N$ in a D-dimensional space, which are drawn independently from an unknown probability distribution Q^*. Then, we want to infer the distribution within the framework of on-line learning. In general, no finite number of example data allows us to determine the distribution uniquely. Therefore, we have to adopt a probabilistic description, which is given by Bayes's rule as the probability of a distribution Q [1].

$$P[Q|P_1,\ldots,P_N]=\frac{P[P_1,\ldots,P_N|Q]\mathcal{P}[Q]}{P(P_1,\ldots,P_N)},$$
$$=\frac{Q(P_1)\cdots Q(P_N)\mathcal{P}[Q]}{\int \mathcal{D}Q Q(P_1)\cdots Q(P_N)\mathcal{P}[Q]}, \tag{1}$$

where $\mathcal{P}[Q]$ denotes a ''prior distribution.'' It is a probability density in the space of possible distributions and encodes our *a priori* knowledge of the function Q.

A. One-dimensional case

First, we start from the case in one-dimensional space. To make our learning problem well posed, we must not introduce too many degrees of freedom to be determined. Therefore, the prior distribution is chosen to define a scalar field theory with a normalization constraint [1] so that it may suppress short wavelength modes of the probability function Q enough. Furthermore, reparametrization invariance requires us to couple the scalar field to one-dimensional gravity [4]. Although the gravity may seem to introduce another local degree of freedom, it causes no problem since it is eliminated through the gauge fixing procedure. Introducing a scalar field $-\infty<\phi<\infty$ and a metric h in one-dimensional space, we write the probability function as $Q=\sqrt{h}e^{-\phi}/l$. We should note that the length of a parameter l can be included in a scalar density $\sqrt{h}$. Therefore, l may be fixed to be of unit length in numerical studies, and only plays a role to ensure the consistency of dimensions. As a result, we can explicitly write down the prior distribution in a coordinate x as

$$\mathcal{P}[Q]=\frac{1}{Z_0}\exp\left[-\frac{l}{2}\int dx\sqrt{h}^{-1}(\partial_x\phi)^2 - \frac{1}{l}\int dx\sqrt{h}F(\phi)\right]\delta\left[\frac{1}{l}\int dx\sqrt{h}e^{-\phi(x)}-1\right]. \tag{2}$$

 ©2001 The American Physical Society

Here, we have introduced an unknown scalar function $F(\phi)$, which is important in the on-line learning scheme and determined by the requirement of renormalizability [2]. The δ function gives the constraint to normalize the probability distribution, and the factor Z_0 is a normalization constant of the prior distribution.

From the prior distribution (2) and Bayes' rule (1), we find the partition function to be $Z=(1/Z_0)\int(d\lambda/2\pi)\int \mathcal{D}h\,\mathcal{D}\phi\,\exp[-(1/g)S(h,\phi,\lambda)]$, where

$$S=\frac{l}{2}\int dx\sqrt{h}^{-1}(\partial_x\phi)^2+\frac{1}{l}\int dx\sqrt{h}F(\phi) +i\lambda\left[\frac{1}{l}\int dx\sqrt{h}e^{-\phi}-1\right]+N\int dx P_N[\phi-\ln\sqrt{h}]. \tag{3}$$

The constant g is introduced to count fluctuation effects, and is set to 1 in numerical analysis. P_N is a scalar density of sample data, given by $P_N(x)=(1/N)\Sigma_{i=1}^N\delta(x-x_i)$.

For further analysis, we consider the asymptotic situation $N\gg 1$ so that we may regard the function $P_N(x)$ as continuous and differentiable. Then, we have to note that the partition function Z is not well defined because of the infinite volume of reparametrization symmetry group. Therefore, we need to divide the function Z by the volume. This can be easily done through the gauge fixing procedure, in which we now pick up the metric h satisfying the condition: $h(x)=(dy/dx)^2$. This gauge fixing condition corresponds to the ''conformal gauge'' in one-dimensional space. Although the gauge fixing condition eliminates a local degree of freedom of the metric, we also have ghost fields through the gauge fixing procedure. However, they decouple to other fields, so that we do not have to consider them in the following. Thus, we can leave one local degree of freedom ϕ, which is appropriate to learning problems.

The gauge condition leads us to observe the system in a reparametrization-invariant coordinate: $y(x)=\int^x\sqrt{h(s)}ds$ [4]. This coordinate is equivalent to the proper time in relativity theories [5]. In the invariant coordinate y, we can rewrite the action (3) in terms of invariant quantities

$$S=\frac{l}{2}\int dy(\partial_y\phi)^2+\frac{1}{l}\int dy\,F(\phi(y)) +i\lambda\left[\frac{1}{l}\int dy e^{-\phi(y)}-1\right]+N\int dy\,P_N(y)\phi(y) -N\int dx\,P_N(x)\ln\sqrt{h(x)}. \tag{4}$$

This form of the action explicitly shows that we obtain only fluctuation effects with reparametrization invariance. Therefore, this ensures that we can choose the function $F(\phi)$, which keeps the invariance of the action.

Now, we evaluate the functional integration of $\phi(y)$ around classical solutions. However, we note that the ϕ field part of the action (4) is identical to the one in the noncovariant case [2]. Therefore, we can apply our previous result of the leading correction to the classical action $S(\hat\phi,\hat h,\hat\lambda)+g(\sqrt{N}/2l)\int dy\,e^{-(1/2)\hat\phi}$, where the classical solutions $\hat\phi(y)$ and $\hat\lambda$ are defined by the equations

$$-l^2\partial_y^2\hat\phi+F'(\hat\phi)-i\hat\lambda e^{-\hat\phi}=-NlP_N(y), \tag{5}$$

$$\frac{1}{l}\int dy\,e^{-\hat\phi(y)}=1. \tag{6}$$

As a result, we introduce a parameter k_N and determine the unknown function F so as to make the action (4) renormalizable.

$$F(\phi)=k_0e^{-(1/2)\phi},\quad k_0=k_N-\frac{g}{2}\sqrt{N}. \tag{7}$$

Then, applying the constraint (6) to the integration of Eq. (5), we find that $(i\hat\lambda)_N=N-k_N\int dy\,e^{-(1/2)\hat\phi_N}/2l$. Hereafter, we explicitly attach the number N of the examples we have already observed.

In order to derive the on-line learning algorithm for the expectation value $\langle\phi\rangle$ of the field ϕ, we note that $\langle\phi\rangle\simeq\hat\phi$ in our approximation level. Then, the variation of Eq. (5) gives the on-line learning algorithm for $\langle\phi\rangle$ in the coordinate y, which is identical to the one in the noncovariant formulation and is proved to be optimal [2],

$$\Delta\langle\phi_N(y)\rangle\simeq\frac{1}{g}\int dy'\,G_N(y,y')\left[-\delta(y'-y_{N+1}) +\frac{\Delta(i\hat\lambda)_N}{l}e^{-\langle\phi_N(y')\rangle}+\frac{1}{2l}\frac{dk_N}{dN}e^{-(1/2)\langle\phi_N(y')\rangle}\right], \tag{8}$$

where $dk_N/dN=g/4\sqrt{N}$ is the *renormalization group equation* of the parameter k_N. The change $\Delta(i\hat\lambda)_N\simeq 1-k_Ne^{(1/2)\langle\phi_N(y_{N+1})\rangle}/4N$ is derived from the variation of the parameter $(i\hat\lambda)_N$. The Green's function $G_N(y,y')$ is the learning rate with $\xi_N(y)$ as local bin size.

$$G_N(y,y')\simeq\frac{g}{2l}\sqrt{\xi_N(y)\xi_N(y')}\exp\left[-\int_{\min(y,y')}^{\max(y,y')}\frac{ds}{\xi_N(s)}\right], \tag{9}$$

$$\xi_N(y)=l\left[(i\hat\lambda)_Ne^{-\langle\phi_N(y)\rangle}+\frac{k_N}{4}e^{-(1/2)\langle\phi_N(y)\rangle}\right]^{-1/2}. \tag{10}$$

Finally, we will derive the relation between x and y coordinates. We obtain the following equation from the variation of the action (4) with respect to the function $y(x)$,

$$\frac{d}{dx}\left[\frac{P_N(x)}{y'(x)}\right]=0. \tag{11}$$

Equation (11) is easily integrated to give $y(x)=(1/N)\Sigma_{i=1}^N\theta(x-x_i)$. Here, we have fixed a global degree of freedom of the metric $h(x)$ so that the function $y(x)$ may

transform the interval $[x_-,x_+]$ to [0,1], where x_- and x_+ are the coordinates of the lower and the upper ends of the interval we observe. In on-line learning scheme, we may construct the function $y(x)$ iteratively,

$$y_{N+1}(x)\simeq\left(1-\frac{1}{N}\right)y_N(x)+\frac{1}{N}\theta(x-x_{N+1}). \tag{12}$$

From the construction of the coordinate y, we easily see the reason why we can obtain a reparametrization-invariant algorithm (8). The coordinate y labels the data points, observed in x coordinate, according to their orders from one end x_- of the interval. Therefore, equivalent distributions under reparametrizations in x coordinate are described identically to each other in the y coordinate. Thus, we have found that the invariant coordinate y plays an essential role in the reparametrization-invariant algorithm.

B. General *D*-dimensional case

Next, we will extend the previous discussion to the general D-dimensional case. For reparametrization invariance, we may construct a prior distribution from various curvatures, which introduce a metric as local degrees of freedom. However, a metric tensor $h_{\mu\nu}(x)$ has $D(D+1)/2$ local degrees of freedom in D-dimensional space. It is one local degree of freedom that we require for learning problems. Therefore in $D>2$, we have redundant degrees of freedom even after the gauge fixing procedure, which eliminates D local degrees of freedom of the reparametrization symmetry group. This redundancy can be naturally resolved when we adopt a conformal flat metric [4], for which there exists a coordinate transformation $y^\mu(x)$ satisfying

$$h_{\mu\nu}(x)=\frac{\partial y^\rho}{\partial x^\mu}\frac{\partial y^\sigma}{\partial x^\nu}e^{-(2/D)\phi(y(x))}\delta_{\rho\sigma}. \tag{13}$$

Here, only one local degree of freedom of the metric $\phi(y)$ is called a *conformal mode*. Such a conformal gauge y^μ is possible in $D>3$ only when we impose the vanishing of Weyl curvature $W_{\mu\nu\rho\sigma}$, while it is always possible in two-dimensional space. As will be explicitly constructed later, the coordinate y^μ is reparametrization invariant. This ensures that we obtain reparametrization-invariant fluctuation effects, which can be canceled out by appropriate counter terms.

Since only a conformal mode ϕ survives under the gauge fixing condition (13), we should define the probability distribution to be $Q=\sqrt{h}/l^D=\det(\partial y^\mu/\partial x^\nu)e^{-\phi}/l^D$. Then, the prior distribution in the gauge (13) has to penalize large gradients of the conformal mode. It is possible to construct various types of prior distributions. However, for the asymptotic situation $N\gg 1$, which is equivalent to the ultraviolet limit, it is enough to take the square of a curvature with an appropriate number of derivatives. We can see that the number of derivatives, rather than the details of curvatures, is crucial for the performance of an algorithm in the asymptotic case, since the number of derivatives determines the cutoff scale of short wavelength modes of the conformal mode. This is also seen from our previous result that the optimal error decay is realized when we choose our hypothesis function from the C^{2D} class [2]. From these reasons, we will limit ourselves to the asymptotic case $N\gg 1$ in the following.

As a result, we choose the following prior distribution so that the conformal mode ϕ may not give ultraviolet divergences $(2\alpha>D)$:

$$P[Q]=\frac{1}{Z_0}\exp\left[-\frac{l^{2\alpha-D}}{2}\int d^Dx\sqrt{h}(\nabla_x^{\alpha-2}R)^2 - \frac{1}{l^D}\int d^Dx\sqrt{h}F(h_{\mu\nu})\right]\delta\left[\frac{1}{l^D}\int d^Dx\sqrt{h}-1\right] \times(\text{constraint for conformal flatness}). \tag{14}$$

Hereafter, we use the notation $\nabla_x^\alpha\equiv\nabla_{x\mu_1}\cdots\nabla_{x\mu_\alpha}$, and mean by the repeated indices $\mu_1,\ldots,\mu_\alpha$ the sums from 1 to the dimension D. We have introduced an unknown scalar function $F(h_{\mu\nu})$, which will be determined later so that we may absorb fluctuation effects by renormalization. The δ function gives the constraint to normalize the probability distribution, and the factor Z_0 is a normalization constant. Also, the ''constraint for conformal flatness'' imposes the condition, which limits the metrics to conformal flat ones. This is given by $\Pi_x\delta(W_{\mu\nu\rho\sigma}(x))$ in $D>3$, while it is needless in two dimensions.

Putting the prior distribution (14) into Bayes' rule (1), we obtain the partition function as $Z=(1/Z_0)\int(d\lambda/2\pi)\int\mathcal{D}h_{\mu\nu}\mathcal{D}B\exp[-(1/g)S(h,B,\lambda)]$,

$$S=\frac{l^{2\alpha-D}}{2}\int d^Dx\sqrt{h}(\nabla_x^{\alpha-2}R)^2+\frac{1}{l^D}\int d^Dx\sqrt{h}F(h_{\mu\nu}) + i\lambda\left[\frac{1}{l^D}\int d^Dx\sqrt{h}-1\right]-N\int d^Dx\,P_N\ln\sqrt{h} + (\text{constraint term}). \tag{15}$$

Here, the constant g is introduced to count fluctuation effects and is set to 1 in numerical analysis. $B(x)$ is an auxiliary field to exponentiate the constraint for conformal flatness, and P_N is the distribution of sample data defined to be $P_N(x)=(1/N)\Sigma_{i=1}^N\delta^D(x-x_i)$.

In the asymptotic case $N\gg 1$, the distribution $P_N(x)$ becomes continuous and differentiable. Then, the action (15) has reparametrization symmetry, and the infinite volume of the symmetry group makes the partition function Z ill defined. Again, we have to divide the partition function Z by the infinite volume of the group. We can gauge fixing by putting the gauge condition (13) into the partition function and adding appropriate ghost terms to the action (15). We have found the asymptotic form of the action (15) in the conformal gauge y^μ [5],

$$\left[\frac{2(D-1)}{D}\right]^2 \frac{l^{2\alpha-D}}{2}\int d^D y\, e^{(4/D-1)\phi}(\partial_y^{\alpha-2}\Delta\phi)^2 + \frac{1}{l^D}\int d^D y\, F(\phi) + i\lambda\left[\frac{1}{l^D}\int d^D y\, e^{-\phi}-1\right] + N\int d^D y\, P_N\phi - N\int d^D x\, P_N \ln\det\left(\frac{\partial y^\mu}{\partial x^\nu}\right) + (\text{constraint and ghost terms}). \tag{16}$$

We expand the action (16) around classical solutions and perform the functional integrations. In the asymptotic situation $N\gg 1$, the contribution from ghost fields is negligible compared with that from the conformal mode. The classical solutions $\hat{\phi}(y)$ and $\hat{\lambda}$ are defined by the following equations:

$$\left[\frac{2(D-1)}{D}\right]^2 (-1)^\alpha l^{2\alpha}\Delta^\alpha\hat{\phi} + F'(\hat{\phi}) - i\hat{\lambda}e^{-\hat{\phi}} = -Nl^D P_N, \tag{17}$$

$$\frac{1}{l^D}\int d^D y\, e^{-\hat{\phi}(y)} = 1. \tag{18}$$

These are the same equations as in the noncovariant case except for the factor $[2(D-1)/D]^2$. Therefore in the conformal gauge y^μ, we can also obtain equivalent results to noncovariant ones. If we apply the normalization condition (18) to the integration of Eq. (17), we can find that $i\hat{\lambda}=N+\int d^D y\, F'(\hat{\phi}(y))/l^D$, which determines the mass of the conformal mode ϕ.

It is straightforward to perform the functional integration of $\phi(y)$. The leading corrections to the classical solutions are extracted from the integration up to quadratic terms, which is given by the ratio of functional determinants. We can evaluate it in a standard way [6]. As a result, fluctuation effects are found to give the action (16) the correction: $gRN^{(D/2\alpha)}\int d^D y\, e^{-(D/2\alpha)\phi}/l^D$, where the constant R is defined as $R=[D/2(D-1)]^{D/\alpha}/(4\pi)^{D/2-1}4D\Gamma(D/2)\sin(D/2\alpha)\pi$.

Now that we have obtained a local correction term, we can determine the unknown function $F(\hat{\phi})$. As is easily seen, a counter term can be produced only through the following definition of the function $F(\hat{\phi})$:

$$F(\phi)=k_0 e^{-(D/2\alpha)\phi}, \tag{19}$$

$$k_0 = k_N - gRN^{D/2\alpha}. \tag{20}$$

Here, we have introduced a parameter k_N. Since a bare parameter k_0 does not depend on the scale N, we find the *renormalization group equation*, which describes the scaling behavior of the parameter k_N.

$$\frac{dk_N}{dN} = g\frac{DR}{2\alpha}\frac{1}{N^{1-D/2\alpha}}. \tag{21}$$

Thus, we can obtain the scaling form of the parameter to be $k_N \simeq gRN^{D/2\alpha}$.

An on-line learning algorithm is how to change our hypothesis about an unknown distribution after receiving a new example. It is given by the recursion relation of the expectation value $\langle\phi(y)\rangle$ of the field $\phi(y)$. In our approximation level, the expectation value is just the classical solution $\hat{\phi}(y)$. Therefore, an on-line learning algorithm is directly obtained from the variation of Eq. (17),

$$\Delta\langle\phi_N(y)\rangle \simeq \frac{1}{g}\int d^D y'\, G_N(y,y')\left[-\delta^D(y'-y_{N+1}) + \frac{\Delta(i\hat{\lambda})_N}{l^D}e^{-\langle\phi_N(y')\rangle} + \frac{D}{2\alpha l^D}\frac{dk_N}{dN}e^{-(D/2\alpha)\langle\phi_N(y')\rangle}\right], \tag{22}$$

which is proved to give optimal performance only when we choose our hypothesis functions from the C^{2D} class, i.e., $\alpha=D$ [2]. Hereafter, we will explicitly attach the number of examples N to denote the scale. The learning rate $G_N(y,y')$ is a Green's function, and defines a local bin size $\xi_N(y)$,

$$G_N(y,y') \simeq \frac{g}{l^{2\alpha-D}}\left[\frac{D}{2(D-1)}\right]^2 \frac{(-1)^{\alpha-1}}{(2\pi)^{D/2}\alpha r^{D/2-1}} \times \sum_{n=0}^{\alpha-1}(\gamma_n\xi_N^{-1})^{D/2-2\alpha+1}K_{D/2-1}(\gamma_n\xi_N^{-1}r), \tag{23}$$

$$\xi_N(y) = l\left[\frac{D}{2(D-1)}\right]^{-1/\alpha}\left[(i\hat{\lambda})_N e^{-\langle\phi_N(y)\rangle} + \left(\frac{D}{2\alpha}\right)^2 k_N e^{-(D/2\alpha)\langle\phi_N(y)\rangle}\right]^{-1/2\alpha}. \tag{24}$$

Here, r is the distance between the points y and y', and the function $K_{D/2-1}$ is a modified Bessel function of the second

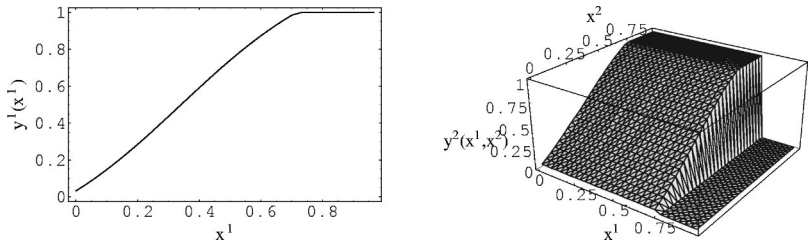

FIG. 1. The coordinate system $\{y^\mu\}$ in learning a Gaussian distribution $Q^*(x)=\exp[-32\{(x^1-3/8)^2+(x^2-3/8)^2\}/9]/Z$ $(0\leqslant x^1,x^2\leqslant 3/4)$, where Z is a normalization constant. We have performed the algorithm (27) with the region divided into 30×30 pieces.

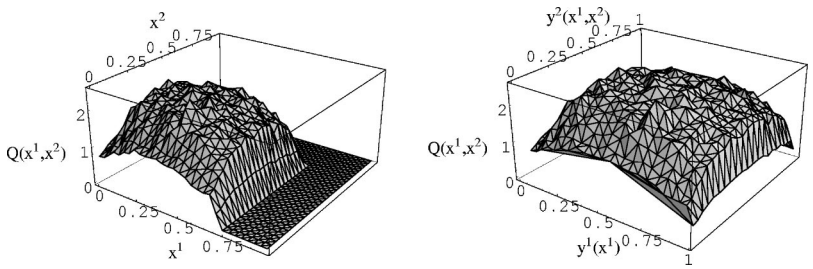

FIG. 2. The numerical result of learning the Gaussian distribution $Q^*(x)=\exp[-32\{(x^1-3/8)^2+(x^2-3/8)^2\}/9]/Z$ $(0\leqslant x^1,x^2\leqslant 3/4)$, where Z is a normalization constant. The left and the right graphs show the result in $\{x^\mu\}$ coordinate system and the one in $\{y^\mu\}$ coordinate system, respectively. We have performed the algorithm (22) for 100 000 data points with the region divided into 30×30 pieces.

kind. γ_n is defined to be $\gamma_n\equiv\exp[i(2n+1)\pi/2\alpha]e^{-i\pi/2}$. The Green's function is real and regular for any $r\geqslant 0$. Also, we have found that the change $\Delta(i\hat{\lambda})_N$ is given by $\Delta(i\hat{\lambda})_N\simeq 1-D^2k_Ne^{(1-D/2\alpha)\langle\phi_N(y_{N+1})\rangle}/4\alpha^2N$, which is the variation of the parameter $(i\hat{\lambda})_N=N-Dk_N\int d^Dy\, e^{-(D/2\alpha)\langle\phi_N\rangle}/2\alpha l^D$.

Next, we will discuss the transformation from the coordinate system $\{x^\mu\}$ to $\{y^\mu\}$. It is obtained from the variation of the action (16) with respect to the coordinate y^μ,

$$\frac{\partial}{\partial x^\nu}\left[\frac{P_N(x)}{\partial_\nu y^\mu(x)}\right]=0. \tag{25}$$

This equation is solved by factorizing the distribution $P_N(x)$ as $P_N=\Pi_{i=1}^D P_N^i$ [4],

$$P_N^1(x^1)=\int\prod_{i=2}^{D}dx^i\,P_N(x),$$

$$P_N^2(x^2|x^1)P_N^1(x^1)=\int\prod_{i=3}^{D}dx^i\,P_N(x),$$

$$\cdots$$

$$P_N^D(x^D|x^1,\ldots,x^{D-1})\prod_{i=1}^{D-1}P_N^i=P_N(x). \tag{26}$$

From these distributions, we can construct the conformal gauge as $y_N^\mu(x)=\int_{x_-^\mu}^{x^\mu}dx^\mu\,P_N^\mu$, which transforms the interval $[x_-^\mu,x_+^\mu]$ to $[0,1]$. Here, x_-^μ and x_+^μ are the minimum and the maximum values of the μ coordinate of the region we observe. In the on-line learning scheme, the function $y_N^\mu(x)$ is obtained iteratively as

$$y_{N+1}^1(x)\simeq\left(1-\frac{1}{N}\right)y_N^1(x)+\frac{1}{N}\theta(x^1-x_{N+1}^1),$$

$$y_{N+1}^2(x)\simeq\left[1-\frac{\delta(x^1-x_{N+1}^1)}{NP_{N+1}^1(x_{N+1}^1)}\right]y_N^2(x)+\frac{\delta(x^1-x_{N+1}^1)}{NP_{N+1}^1(x_{N+1}^1)}\times\theta(x^2-x_{N+1}^2),\ \ldots\ . \tag{27}$$

Here, $\delta(x^1-x_{N+1}^1)/NP_{N+1}^1(x_{N+1}^1)$ is the inverse of the number of the data, of which x^1 coordinates are equal to x_{N+1}^1. So, the second equation of Eq. (27) has the same form as the first when $x^1=x_{N+1}^1$, otherwise it gives $y_{N+1}^2=y_N^2$. The conformal gauge y^μ puts labels to the example data, observed in the x^μ coordinate, according to their orders from the end x_-^μ of the interval $[x_-^\mu,x_+^\mu]$. It is clear that this coordinate system has reparametrization invariance. Therefore in the coordinate system $\{y^\mu\}$, we can identify the distributions that are equivalent under reparametrizations.

Figures 1 to 4 show the result of the numerical simulation of the algorithms (22) and (27). They are applied to the learning of two-dimensional Gaussian distributions, which are transformed from one to another by a scale transformation and a shift (the graphs on the left-hand sides of Figs. 2 and 4). Although they look different in the $\{x^\mu\}$ coordinate system, we can recognize in the $\{y^\mu\}$ coordinate system (Figs. 1 and 3) that they are equivalent distributions to each other (the graphs on the right-hand sides of Figs. 2 and 4).

III. CONCLUSIONS AND DISCUSSIONS

In this paper, we have discussed the on-line learning of probability distributions in a reparametrization-covariant framework. Reparametrization covariance is fundamental from the information theoretical point of view, since it is an intrinsic property of ''information'' that it does not depend on a specific coordinate system to observe it. We have ob-

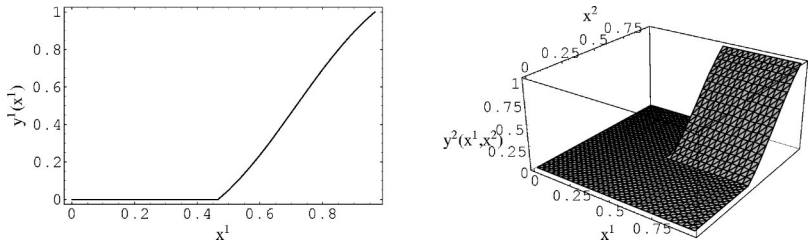

FIG. 3. The coordinate system $\{y^\mu\}$ in learning a Gaussian distribution $Q^*(x)=\exp[-8\{(x^1-3/4)^2+(x^2-3/4)^2\}/Z$ $(1/2\leqslant x^1,x^2\leqslant 1)$, where Z is a normalization constant. We have performed the algorithm (27) with the region divided into 30×30 pieces.

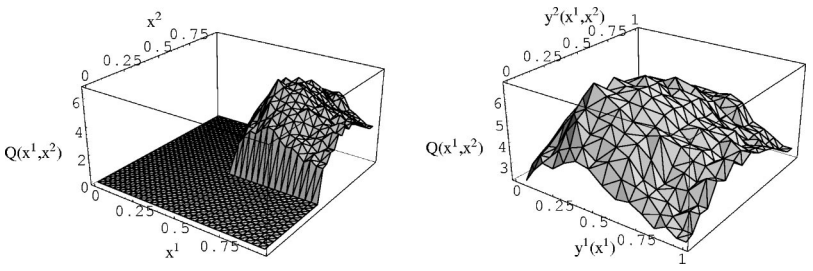

FIG. 4. The numerical result of learning the Gaussian distribution $Q^*(x)=\exp[-8\{(x^1-3/4)^2+(x^2-3/4)^2\}/Z$ $(1/2\leqslant x^1,x^2\leqslant 1)$, where Z is a normalization constant. The left and the right graphs show the result in $\{x^\mu\}$ coordinate system and $\{y^\mu\}$ coordinate system, respectively. We have performed the algorithm (22) for 100 000 data points with the region divided into 30×30 pieces.

tained an on-line learning algorithm, which is independent of a specific coordinate system to observe example data. In the following, we will consider the properties of the reparametrization-invariant algorithm.

First, let us consider the performance of the algorithm. For simplicity, we will limit ourselves to the one-dimensional case. The quadratic error of the algorithm is found to decay as $1/N$ in a noncovariant framework [2],

$$\left\langle \left\langle \int dx\, Q^*(x)\epsilon_N(x)^2 \right\rangle \right\rangle = \frac{V_p V_x}{2\pi}\frac{1}{N}, \tag{28}$$

where coefficients V_p and V_x denote the volumes of momentum and coordinate space, respectively. The coefficient V_p comes from the relation $\delta(x=0)\simeq V_p/2\pi$. Here, we have to note that $\delta(x=0)$ is not equal to infinity in practice. We are not able to use infinite momentum or infinitesimal wavelength for observation, since we have the physical limit of a maximum resolution. Therefore, we should replace $\delta(x=0)$ with $V_p/2\pi$ and consider its precise meaning. When we have a length a of lattice spacing as a maximum resolution length, a half wavelength in the space is able to take a discrete value from a to V_x at the interval of a. Then, V_p is proportional to the number of bins V_x/a and simply reflects the physical complexity of the system we use. This result is equivalent to the universal asymptotic behavior, which is well known in neural network models [7,8].

In the covariant formulation, the conformal gauge plays a crucial role for reparametrization invariance. Once we describe distributions in invariant coordinates, we may follow the algorithm in the noncovariant case. Therefore, we can also find Eq. (28), not in a coordinate system $\{x^\mu\}$, but in an invariant one $\{y^\mu\}$. Then, the coefficient V_p takes into account the local properties of distributions as well as the physical complexity of the system, since the coordinate system $\{y^\mu\}$ locally changes the lattice spacing a according to the integral of distributions. We can see that V_p in the covariant case is smaller that in the non-covariant one, in general. This is easily found in learning a very local constant distribution, for example. As a result, we can expect better performance than in noncovariant case.

Second, the invariant algorithm makes the quadratic error uniformly convergent [4]. A remarkable result is that our inference makes sense even in the regions where we can observe a small number of examples. This is traced to the fact that our geometry to observe data is locally controlled by the metric, which is the local counterpart of the *a priori* length scale l.

On the other hand, covariant formulation has an advantage over the noncovariant one also in practical applications. As is seen in numerical results, we can apply the algorithm to pattern recognition problems such as visual and speech information processes. This is because an invariant coordinate system $\{y^\mu\}$ enables us to identify the distributions that transform one another by reparametrization.

ACKNOWLEDGMENTS

The author would like to thank Y. Kitazawa and Y. Watabiki for their valuable comments and discussions. This work was partially supported by the Grant-in-Aid for Scientific Research No. 10750056 from the Ministry of Education, Science and Culture, Japan.

[1] W. Bialek, C.G. Callan, and S.P. Strong, Phys. Rev. Lett. **77**, 4693 (1996).
[2] T. Aida, Phys. Rev. Lett. **83**, 3554 (1999).
[3] See, for example, N. Murata, S. Yoshizawa, and S. Amari, IEEE Trans. Neural Netw. **5**, 865 (1994).
[4] V. Periwal, Phys. Rev. Lett. **78**, 4671 (1997); V. Periwal, Nucl. Phys. B **554**, 719 (1999).
[5] R.M. Wald, *General Relativity* (University of Chicago Press, Chicago, 1984), p. 445.
[6] S. Coleman, *Aspects of Symmetry* (Cambridge University Press, Cambridge, 1975), p. 340.
[7] H. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992); S. Amari and N. Murata, Neural Comput. **5**, 140 (1993); M. Opper and D. Haussler, Phys. Rev. Lett. **75**, 3772 (1995).
[8] M. Opper, Phys. Rev. Lett. **77**, 4671 (1996).