entropy

MDPI

Article

Rate-Distortion Bounds for Kernel-Based Distortion Measures †

Kazuho Watanabe

Department of Computer Science and Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka Tempaku-cho Toyohashi, Aichi 441-8580, Japan; wkazuho@cs.tut.ac.jp; Tel.: +81-532-44-6893

† This paper is an extended version of my papers published in the Eighth Workshop on Information Theoretic Methods in Science and Engineering, Copenhagen, Denmark, 24–26 June 2015 and the IEEE International Symposium on Information Theory, Aachen, Germany, 25–30 June 2017.

Received: 9 May 2017; Accepted: 2 July 2017; Published: 5 July 2017

Abstract: Kernel methods have been used for turning linear learning algorithms into nonlinear ones. These nonlinear algorithms measure distances between data points by the distance in the kernel-induced feature space. In lossy data compression, the optimal tradeoff between the number of quantized points and the incurred distortion is characterized by the rate-distortion function. However, the rate-distortion functions associated with distortion measures involving kernel feature mapping have yet to be analyzed. We consider two reconstruction schemes, reconstruction in input space and reconstruction in feature space, and provide bounds to the rate-distortion functions for these schemes. Comparison of the derived bounds to the quantizer performance obtained by the kernel K-means method suggests that the rate-distortion bounds for input space and feature space reconstructions are informative at low and high distortion levels, respectively.

Keywords: kernel methods; rate-distortion function; kernel K-means; preimaging

1. Introduction

Kernel methods have been widely used for nonlinear learning problems combined with linear learning algorithms such as the support vector machine and the principal component analysis [1]. By the so-called kernel trick, kernel-based methods can use linear learning methods in the kernel-induced feature space without explicitly computing the high-dimensional feature mapping. Kernel-based methods measure the dissimilarity between data points by the distance in the feature space, which, in input space, corresponds to a distance measure involving the feature mapping [2]. If a kernel-based learning method is used as a lossy source coding scheme, its optimal rate-distortion tradeoff is indicated by the rate-distortion function associated with the distortion measure defined by the kernel feature map [3]. Successful applications of kernel methods in learning problems and flexibility to create various distance measures suggest that kernel-based distortion measures can be suitable for certain lossy compression problems. However, the rate-distortion function of such a distortion measure has yet to be evaluated analytically. Although there are several kernel-based approaches to vector quantization [4,5], their rate-distortion tradeoffs are still unknown.

In this paper, we derive bounds for the rate-distortion functions for kernel-based distortion measures. We consider two schemes to reconstruct inputs in lossy coding methods. One is to obtain a reconstruction in the original input space. Since kernel methods usually yield results of learning by the linear combination of vectors in feature space, we need an additional step to obtain the reconstruction in input space, such as preimaging [6]. The other is to consider the linear combination of feature vectors as the reconstruction and measure the distortion in the feature space directly. We formulate the two reconstruction schemes (Sections 3.1 and 3.2), and prove that the rate-distortion function of input space reconstruction provides an upper bound of that of feature space reconstruction (Section 3.3). We derive

lower and upper bounds to the rate-distortion function of input space reconstruction, which are computable only by *one*-dimensional numerical integrations in the case of translation invariant and isotropic kernel functions (Sections 4.1 and 4.2). We also provide an upper bound to the rate-distortion function of feature space reconstruction for general positive definite kernel functions (Section 4.4). In the usual applications of kernel-based quantization algorithms, one fixes the rate by determining the number of quantized points, and minimizes the average distortion for training data. The distortion-rate function, which is the inverse function of the rate-distortion function, shows the minimum achievable expected distortion (or distortion for test data) at the fixed rate. The derived bounds approximately characterize such optimal tradeoffs between the rate and expected distortion.

Furthermore, we design a vector quantizer using the kernel K-means method and compare its performance with the derived rate-distortion bounds (Section 5). We also compute the preimages of the quantized points in feature space to investigate the performance of the quantizer in input space. It is suggested through the experiments using synthetic and image data that the rate-distortion bounds of reconstruction in input space are accurate at low distortion levels while the upper bound for reconstruction in feature space is informative at high distortion levels.

2. Rate-Distortion Function

Let X and Y be random variables of input and reconstruction taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. For the non-negative distortion measure between x and y, $d(x,y)$, the rate-distortion function $R(D)$ of the source $X \sim p(x)$ is defined by

$$R(D) = \inf_{q(y|x):E[d(X,Y)]\leq D} I(q), \tag{1}$$

where $I(q) = I(X;Y)$ is the mutual information and E denotes the expectation with respect to $q(y|x)p(x)$. $R(D)$ shows the minimum achievable rate R under the given distortion measure d [3,7]. The distortion-rate function is the inverse function of the rate-distortion function and denoted by $D(R)$.

If the conditional distributions $q_s(y|x)$ achieve the minimum of the following Lagrange functional parameterized by $s \geq 0$,

$$L(q) = I(q) + s\left(E[d(X,Y)] - D\right),$$

then, the rate-distortion function is parametrically given by

$$\begin{aligned} R(D_s) &= I(q_s), \\ D_s &= \int q_s(y|x)p(x)d(x,y)dxdy. \end{aligned}$$

The parameter s corresponds to the (negated) slope of the tangent of $R(D)$ at $(D_s, R(D_s))$ and hence is referred to as the slope parameter [3]. Alternatively, if there exists a marginal reconstruction density $q_s(y)$ that minimizes the functional,

$$F(q) = -\frac{1}{s}E\left[\log \int e^{-sd(X,y)}q(y)dy\right],$$

then the optimal conditional reconstruction distributions are given by

$$q_s(y|x) = \frac{e^{-sd(x,y)}q_s(y)}{\int e^{-sd(x,y)}q_s(y)dy} \tag{2}$$

(see, for example, [3,8]).

From the properties of the rate-distortion function $R(D)$, we know that $R(D) > 0$ for $0 < D < D_{\max}$, where

$$D_{\max} = \inf_y \int p(x) d(x,y) dx, \tag{3}$$

and $R(D) = 0$ for $D \geq D_{\max}$ [3] (p. 90). Hence, $D_{\max} = \lim_{R \to 0} D(R)$.

3. Kernel-Based Distortion Measures

In kernel-based learning methods, data points in input space $\mathcal{X}$ are mapped into some high-dimensional feature space H by a feature mapping ϕ. Then, the similarity between the two points x and y in $\mathcal{X}$ is measured by the inner product $\langle \phi(x), \phi(y) \rangle$ in H.

The inner product is directly evaluated by a nonlinear function in input space

$$K(x,y) = \langle \phi(x), \phi(y) \rangle, \tag{4}$$

which is called the kernel function. Mercer's theorem ensures that there exists some ϕ such that Equation (4) holds if K is a positive definite kernel [1]. This enables us to avoid explicitly computing the feature map ϕ in the potentially high-dimensional space H, which is called the *kernel trick*. A lot of learning methods that can be expressed by only the inner products between data points have been kernelized [1].

We identify the feature space H with the reproducing kernel Hilbert space (RKHS) associated with the kernel function K by the canonical feature map, $\phi(x) = K(\cdot, x)$ [9] (Lemma 4.19). We assume that the input space $\mathcal{X}$ is a subset of $\mathbb{R}^m$, and the kernel function K is continuous [9] (Lemma 4.29). We focus on the squared norm in feature space as the distortion measure, and consider two reconstruction schemes in the following respective subsections.

3.1. Reconstruction in Input Space

If we restrict ourselves to the reconstruction in input space, that is, the reconstruction $y \in \mathcal{X} \subset \mathbb{R}^m$ is computed for each input $x \in \mathcal{X}$, the distortion measure is naturally defined by

$$\begin{aligned} d_{\text{inp}}(x,y) &= ||\phi(x) - \phi(y)||^2 \\ &= K(x,x) + K(y,y) - 2K(x,y). \end{aligned} \tag{5}$$

Note that the reconstruction $\phi(y)$ of $\phi(x)$ is restricted to the subset of the feature space, $\{\phi(y);\ y \in \mathcal{X}\}$. To obtain a reconstruction in input space, we need a technique such as preimaging [6].

This is a difference distortion measure if and only if the kernel function is translation invariant, that is, $K(x+a, y+a) = K(x,y)$ for any $a \in \mathcal{X}$. In this case, the distortion measure is expressed as

$$d_{\text{inp}}(x,y) = \rho(x-y), \tag{6}$$

where $\rho(z) = 2(C - K(z,0))$ and $C = K(0,0)$. The rate-distortion function (distortion-rate function, resp.) for this distortion measure is denoted by $R_{\text{inp}}(D)$ ($D_{\text{inp}}(R)$, resp.) and the maximum distortion $D_{\max}$ in Equation (3) is denoted by $D_{\max,\text{inp}}$, that is,

$$D_{\max,\text{inp}} = E[K(X,X)] + \inf_y \{K(y,y) - 2E[K(X,y)]\}, \tag{7}$$

which is in the translation invariant case, $D_{\max,\text{inp}} = 2\left(C - \sup_y E[K(X,y)]\right)$.

3.2. Reconstruction in Feature Space

Suppose we have a sample of length n in input space, $S = \{x_1, ..., x_n\}$ so that $\{\phi(x_1), ..., \phi(x_n)\}$ spans a linear subspace in feature space. If we compute the reconstruction by the linear combination $\sum_{i=1}^n \alpha_i \phi(x_i)$ for $\alpha_i \in \mathbb{R}, i = 1, ..., n$, and consider it as the reconstruction in feature space, the distortion can be measured by

$$d_{\mathrm{fea}}(x,\boldsymbol{\alpha}) = d_{\mathrm{fea}}^{[S]}(x,\boldsymbol{\alpha}) = \left\| \phi(x) - \sum_{i=1}^{n} \alpha_i \phi(x_i) \right\|^2 = K(x,x) - 2\boldsymbol{\alpha}^T \boldsymbol{k}(x) + \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha}, \tag{8}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^T \in \mathbb{R}^n$,

$$\boldsymbol{k}(x) = (K(x_1,x), ..., K(x_n,x))^T,$$

and $\boldsymbol{K} = (K(x_i,x_j))_{ij}$ is the Gram matrix. Note that the reconstruction is identified with the coefficients $\boldsymbol{\alpha}$ whose domain is not identical to the input space $\mathcal{X}$. Although the distortion measure d_{fea} depends on the sample S, we omit the dependence in the notation since we consider a fixed design of S for a sufficiently large n. The sample does not have to be distributed according to the source distribution, while it is required to overspread the support of the source.

The rate-distortion function (distortion-rate function, resp.) for this distortion measure is denoted by $R_{\mathrm{fea}}(D)$ ($D_{\mathrm{fea}}(R)$, resp.) and the maximum distortion $D_{\max}$ in Equation (3) is given by

$$D_{\max,\mathrm{fea}} = E[K(X,X)] - E[\boldsymbol{k}(X)]^T \boldsymbol{K}^{-1} E[\boldsymbol{k}(X)], \tag{9}$$

which is derived from the direct minimization of the quadratic function of $\boldsymbol{\alpha}$, $\int d_{\mathrm{fea}}(x,\boldsymbol{\alpha}) p(x) dx$.

3.3. $R_{\mathrm{inp}}(D)$ and $R_{\mathrm{fea}}(D)$

The following theorem claims that $R_{\mathrm{inp}}(D)$ provides an upper bound of $R_{\mathrm{fea}}(D)$ when n is sufficiently large.

Theorem 1. *If the input space $\mathcal{X}$ is bounded, and there exists a conditional density achieving the infimum in the definition of $R_{\mathrm{inp}}(D)$, for any $\varepsilon > 0$, $D \geq \varepsilon$, and sufficiently large n, the following inequality holds:*

$$R_{\mathrm{fea}}(D+\varepsilon) \leq R_{\mathrm{inp}}(D).$$

The proof is given in Appendix A. This theorem shows that the feature space reconstruction gives better rates since a single feature vector $\phi(y)$ can be approximated by a linear combination $\sum_{i=1}^{n} \alpha_i \phi(x_i)$ when n is sufficiently large.

4. Rate-Distortion Bounds

Since the rate-distortion problem (Section 2) is rarely solved in a closed form [8], we derive bounds to $R_{\mathrm{inp}}(D)$ and $R_{\mathrm{fea}}(D)$.

4.1. Lower Bound to $R_{\mathrm{inp}}(D)$

Although the Shannon lower bound to $R(D)$ is defined for difference distortion measures in general [3] (p. 92), it diverges to $-\infty$ for the distortion measure in Equation (6) since $\int e^{-s\rho(z)} dz$ diverges to ∞. Hence, we consider an improved lower bound, which was introduced by [3] (p. 140). Let Q_B be the probability that $\|X\| \leq B$. Then, $R(D)$ is lower-bounded as

$$R(D) \geq Q_B \left\{ h(p_B) - \max_{g \in G_{B,D}} h(g) \right\}, \tag{10}$$

where h denotes the differential entropy,

$$p_B(x) = \frac{1}{Q_B} p(x) u(B - \|x\|), \tag{11}$$

and u is the step function. $G_{B,D}$ is the set of all probability densities $g(\cdot)$ for which $g(x) = 0$ for $\|x\| > B$ and $\int \rho(z) g(z) dz \leq D/Q_B$.

In the case of the distortion measure in Equation (6), the maximum in Equation (10) is explicitly given by

$$g_s(z) = \frac{1}{C_{B,s}} \exp\left(2sK(z,0)\right) u(B - \|z\|), \tag{12}$$

where $C_{B,s} = \int_{\|z\| \le B} e^{2sK(z,0)} dz$ for s related to D by $\int \rho(z) g_s(z) dz = D/Q_B$. Since its differential entropy is

$$h(g_s) = -s \frac{\partial \log C_{B,s}}{\partial s} + \log C_{B,s}, \tag{13}$$

we arrive at the following theorem.

Theorem 2. *The rate distortion function $R_{\mathrm{inp}}(D)$ is parametrically lower-bounded as*

$$\begin{aligned} R_{\mathrm{inp}}(D_s) \ge R_{\mathrm{inp},L}(D_s) &= Q_B \left\{ h(p_B) + s \frac{\partial \log C_{B,s}}{\partial s} - \log C_{B,s} \right\}, \\ D_s &= Q_B \left\{ 2C - \frac{\partial \log C_{B,s}}{\partial s} \right\}. \end{aligned} \tag{14}$$

If we further assume that the kernel function is radial, that is, $K(x,y) = K(x-y,0) = k(\|x-y\|)$ for some function k, the integrations above reduce to *one*-dimensional ones,

$$C_{B,s} = A(m) \int_0^B r^{m-1} e^{2sk(r)} dr,$$

and

$$\begin{aligned} \frac{\partial \log C_{B,s}}{\partial s} &= 2 \int_{\|z\| \le B} K(z,0) e^{2sK(z,0)} dz \\ &= 2A(m) \int_0^B r^{m-1} k(r) e^{2sk(r)} dr, \end{aligned} \tag{15}$$

where $A(m) = \frac{m\sqrt{\pi}^m}{\Gamma(m/2+1)}$ is the area of the m-dimensional unit sphere, and Γ is the gamma function.

4.2. Upper Bound to $R_{\mathrm{inp}}(D)$

If d_{inp} in Equation (5) is a difference distortion measure, that is, K is translation invariant, by choosing $q(y|x) = g_s(y-x)$ for the density g_s in Equation (12), the following upper bound is obtained,

$$R_{\mathrm{inp}}(D_s) \le R_{\mathrm{inp},U}(D_s) = h(g_s * p) - h(g_s) \tag{16}$$

$$D_s = 2C - \frac{\partial \log C_{B,s}}{\partial s}, \tag{17}$$

where $h(g_s)$ is given by Equation (13) and $(g_s * p)(y) = \int g_s(y-x) p(x) dx$ is the convolution between g_s and p. This type of upper bound was used to prove the asymptotic tightness of the Shannon lower bound (as $D \to 0$) for a class of general sources and distortion measures [3,10–12]. However, this upper bound requires the evaluation of the differential entropy of the convolution.

The following theorem is derived from the facts that the spherical Gaussian distribution maximizes the entropy under the constraint that $E[\|X\|^2]$ is no greater than a constant, and that $E[\|Y\|^2] = E[\|X\|^2] + E[\|Z\|^2]$ holds for $Y = X + Z \sim g_s * p$.

Theorem 3. *If the kernel function is translation invariant and radial, $K(x,y) = k(\|x-y\|)$, then $R_{\mathrm{inp}}(D)$ is parametrically upper-bounded as*

$$R_{\mathrm{inp}}(D_s) \le R_{\mathrm{inp},G}(D_s) = \frac{m}{2}\log(2\pi e(v_p + v_s)) - h(g_s),$$

where

$$\begin{array}{rcl} v_p & = & \frac{1}{m}\int \|x-\mu\|^2 p(x)dx, \\ \mu & = & \int x p(x)dx, \\ v_s & = & \frac{1}{m}\int \|x\|^2 g_s(x)dx \\ & = & \frac{A(m)}{mC_{B,s}}\int_0^B r^{m+1}e^{2sk(r)}dr, \end{array} \tag{18}$$

and D_s is given by Equation (17) (and Equation (15)).

4.3. Rate-Distortion Dimension

In this section, we evaluate the rate-distortion dimension [13] of the kernel-based distortion measure in Equation (5) to investigate its property. We focus on the radial kernel, $K(x,y) = k(\|x-y\|)$, also in this section, and assume that

$$\lim_{r\to 0}\frac{k(r)-k(0)}{r^\alpha} = -\beta \tag{19}$$

holds for some $\alpha > 0$ and $\beta > 0$. For example, the Gaussian kernel, $k(r) = \exp\left(-\gamma r^2\right)$ $(\gamma > 0)$, satisfies Equation (19) for $\alpha = 2$ and $\beta = \gamma$.

To examine the limit $D \to 0$ of $R_{\mathrm{inp}}(D)$, we consider the asymptotic case of $s \to \infty$. Since $k(r) = k(0) - \beta r^\alpha + o(r^\alpha)$, it follows that

$$\begin{array}{rcl} C_{B,s} & = & A(m)\displaystyle\int_0^B e^{2sk(r)r^{m-1}}dr \\ & = & A(m)e^{2sk(0)}\dfrac{1}{\alpha}\left(\dfrac{1}{s\beta}\right)^{m/\alpha}\left\{\Gamma\left(\dfrac{m}{\alpha}\right)+o(1)\right\}, \end{array}$$

$$\int_0^B 2k(r)e^{2sk(r)r^{m-1}}dr = 2k(0)\frac{C_{B,s}}{A(m)} - 2e^{2sk(0)}\frac{1}{s\alpha}\left(\frac{1}{s\beta}\right)^{1+m/\alpha}\left\{\Gamma\left(1+\frac{m}{\alpha}\right)+o(1)\right\},$$

and

$$\begin{array}{rcl} \dfrac{\partial \log C_{B,s}}{\partial s} & = & \dfrac{\int_0^B 2k(r)e^{2sk(r)r^{m-1}}dr}{\int_0^B e^{2sk(r)r^{m-1}}dr} \\ & = & 2k(0) - \dfrac{m}{s\alpha\beta} + o\left(\dfrac{1}{s}\right). \end{array}$$

Thus, we have from Equations (14) and (17),

$$-\log D_s = \log s + O(1),$$

for both the lower and upper bounds, and from Equation (13),

$$\begin{array}{rcl} h(g_s) = & -\dfrac{m}{\alpha}\log s + O(1) \\ = & \dfrac{m}{\alpha}\log D_s + O(1). \end{array} \tag{20}$$

Since d_{inp} in Equation (5) is a norm squared for a valid RKHS kernel K, the rate-distortion dimension of the source distribution p is defined by [13],

$$\dim_R(p) = \lim_{D\to 0}\frac{R_{\mathrm{inp}}(D)}{-\frac{1}{2}\log D}. \tag{21}$$

From Theorems 2 and 3 and Equation (20), we conclude the following.

Theorem 4. *If the source has a finite differential entropy, positive and finite v_p defined in Equation (18), and a bounded support, that is, there exists a finite $B > 0$ such that $Q_B = 1$ in Equation (11), and the radial kernel, $K(x,y) = k(\|x-y\|)$ satisfies Equation (19) for $\alpha > 0$ and $\beta > 0$, then the rate-distortion dimension Equation (21) of $R_{\mathrm{inp}}(D)$ is given by*

$$\dim_R(p) = \frac{2m}{\alpha}. \tag{22}$$

This theorem shows that the rate-distortion dimension is dependent only on the dimensionality of the input space and independent of the dimensionality of the feature space. In the case of the linear kernel, $K(x,y) = \langle x,y\rangle$, with $\phi(x) = x$, the distortion measure in Equation (5) reduces to the usual squared distortion measure, $\|x-y\|^2$. It can be shown that under norm-based distortion measures including the squared distortion measure, the rate-distortion dimension of a source with an m-dimensional density is m [11,12]. From the preceding theorem, this is also the case for a general radial kernel if the kernel function has the order $\alpha = 2$ as the Gaussian kernel. Expression (22) of the rate-distortion dimension will be examined through a numerical experiment in Section 5.1.

4.4. Upper Bound to $R_{\mathrm{fea}}(D)$

We construct an upper bound to the rate-distortion function $R_{\mathrm{fea}}(D)$. We choose the conditional distribution of the reconstruction by

$$q(\boldsymbol{\alpha}|x) = N(\boldsymbol{\alpha}; \boldsymbol{m}_K(x), \tilde{\boldsymbol{K}}^{-1}/2s), \tag{23}$$

where $\tilde{\boldsymbol{K}} = \boldsymbol{K} + c\boldsymbol{I}$,

$$\boldsymbol{m}_K(x) = \tilde{\boldsymbol{K}}^{-1}\boldsymbol{k}(x),$$

and $N(\cdot; \boldsymbol{m}, \boldsymbol{\Sigma})$ denotes the n-dimensional normal density with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Sigma}$. Here, we have introduced the regularization constant $c \geq 0$ with the $n \times n$ identity matrix $\boldsymbol{I}$. The conditional distribution in Equation (23) is implied by Equation (2) and the approximation $q_s(\boldsymbol{\alpha}) = N(\boldsymbol{\alpha}; \boldsymbol{0}, \boldsymbol{I}/(2sc))$. This reconstruction distribution yields the following upper bound:

$$R_{\mathrm{fea}}(D_s) \leq R_{\mathrm{fea},U}(D_s) = h(M_p) - h(N(\boldsymbol{\alpha}; \boldsymbol{m}_K(x), \tilde{\boldsymbol{K}}^{-1}/2s)), \tag{24}$$

$$D_s = \frac{n - c\mathrm{tr}\{\tilde{\boldsymbol{K}}^{-1}\}}{2s} + D_{\min}(c), \tag{25}$$

where $M_p(\boldsymbol{\alpha}) = \int N(\boldsymbol{\alpha}; \boldsymbol{m}_K(x), \tilde{\boldsymbol{K}}^{-1}/2s)p(x)dx$,

$$h(N(\boldsymbol{\alpha}; \boldsymbol{m}_K(x), \tilde{\boldsymbol{K}}^{-1}/2s)) = \frac{n}{2}\log\left(\frac{\pi e}{s}|\tilde{\boldsymbol{K}}|^{1/n}\right), \tag{26}$$

which is independent of the input x, and

$$D_{\min}(c) = E[K(X,X)] - \mathrm{tr}\{\tilde{\boldsymbol{K}}^{-1}E[\boldsymbol{k}(X)\boldsymbol{k}(X)^T]\} - c\mathrm{tr}\{\tilde{\boldsymbol{K}}^{-1}E[\boldsymbol{k}(X)\boldsymbol{k}(X)^T]\tilde{\boldsymbol{K}}^{-1}\}.$$

If $c = 0$, $D_{\min}$ is the mean of the variance of the prediction by the associated Gaussian process [14].

Further upper-bounding the differential entropy $h(M_p)$ by the Gaussian entropy, we have the following theorem.

Theorem 5. *The rate distortion function $R_{\mathrm{fea}}(D)$ is upper-bounded as*

$$R_{\mathrm{fea}}(D) \leq R_{\mathrm{fea},G}(D) = \frac{1}{2}\log\left|\boldsymbol{I} + \frac{n - c\mathrm{tr}\{\tilde{\boldsymbol{K}}^{-1}\}}{D - D_{\min}(c)}\tilde{\boldsymbol{K}}^{-1}\boldsymbol{C}\right|, \tag{27}$$

where

$$\boldsymbol{C} = E[\boldsymbol{k}(X)\boldsymbol{k}(X)^T] - E[\boldsymbol{k}(X)]E[\boldsymbol{k}(X)]^T. \tag{28}$$

The proof is put in Appendix B. In the simplest case where $\phi(x) = x \in \mathbb{R}^1$, $n = 1$, and the source is the Gaussian, $p(x) = N(x; 0, \sigma^2)$, the upper bound in Equation (27) reduces to

$$R_{\text{fea},G}(D) = \frac{1}{2}\log\left(1 + \frac{\sigma^2}{D}\right),$$

which is an asymptotically (as $D \to 0$) tight upper bound of the well-known rate distortion function for the Gaussian source under the squared distortion measure, $R(D) = \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right)$ [3,7].

5. Experimental Evaluation

We numerically evaluate the rate-distortion bounds obtained in the previous section. Designing a quantizer by the kernel K-means algorithm, we compare its performance with the bounds.

We focus on the case of the Gaussian kernel,

$$K(x, y) = e^{-\gamma\|x-y\|^2} \tag{29}$$

with the kernel parameter $\gamma > 0$.

5.1. Synthetic Data

As a source, we first assumed the uniform distribution on the union of the two regions, $C_1 = \{x \in \mathbb{R}^m; A(m)\|x\|^m \le m/2\}$ and $C_2 = \{x \in \mathbb{R}^m; m^2 \le A(m)\|x\|^m \le m(m+1/2)\}$, where C_1 and C_2 have equal volumes and $C_1 \cup C_2$ has volume 1. This suggests that $B = \left\{\frac{m(m+1/2)}{A(m)}\right\}^{1/m}$ and $Q_B = 1$ in Equation (10) and succeeding equations in Sections 4.1 and 4.2.

We used the trapezoidal rule to compute the *one*-dimensional integrations in the lower bound $R_{\text{inp},L}$ and the upper bound $R_{\text{inp},G}$. We generated i.i.d sample of the size $n = 200$ from the source to compute $\boldsymbol{k}(x)$ and $\boldsymbol{K}$ for $R_{\text{fea},G}$ in Equation (27). Generating another 4000 data points, we approximated the required expectations. We optimized the regularization coefficient c to minimize the upper bound $R_{\text{fea},G}$ for each D.

Using the same data set of the size 4000 as a training data set, we run the kernel K-means algorithm 10 times with random initializations to obtain the minimum distortion for each rate. Varying the number K of quantized points from 2^1 to 2^{10}, for each K, we counted the effective number $\mathtt{K_{eff}}$ of quantized points which have at least one assigned data point and computed the rate by $\log_2 \mathtt{K_{eff}}$ as the quantizer is first order, that is, the block length is one. The kernel parameter γ was chosen so that the clear separation of C_1 and C_2 is obtained when $\mathtt{K} = 2$.

After the training, we computed the distortion and rate for the test data set, by assigning each of 20,000 test data generated from the same source to the nearest quantized points in the feature space.

For each quantized point, we obtained its preimage. That is, if the kth quantized point is expressed as $\sum_{i=1}^n \alpha_{ki}\phi(x_i)$, its preimage is

$$\begin{aligned} y_k &= \operatorname*{argmin}_y \left\| \phi(y) - \sum_{i=1}^n \alpha_{ki}\phi(x_i) \right\|^2 \\ &= \operatorname*{argmax}_y \sum_{i=1}^n \alpha_{ki} K(y, x_i). \end{aligned}$$

We used the mean shift procedure for the maximization, although this procedure only guarantees the convergence to a local maximum [15,16].

The obtained bounds and the quantizer performances are displayed in Figure 1a,b and for $m = 2$ and $m = 10$, respectively, in the forms of distortion-rate functions. The values of $D_{\max}$ in Equations (7) and (9) are also indicated in the figures.

In both dimensions, the upper bound $D_{\mathrm{fea},G}$ is smaller than $D_{\mathrm{inp},G}$ at low rates while the bound is above the quantizer performance. However, the value of $D_{\max,\mathrm{fea}}$ suggests that the bound is informative at low rates. As the rate becomes higher, the lower and upper bounds of the input space reconstruction, $D_{L,\mathrm{inp}}$ and $D_{G,\mathrm{inp}}$, approach each other. In fact, they sandwich the quantizer performance tightly in the *two*-dimensional case, which suggests that the rate-distortion function for the feature space reconstruction, $R_{\mathrm{fea}}(D)$ is close to the rate-distortion function of the input space reconstruction $R_{\mathrm{inp}}(D)$ at high rates.

We see that the quantizer performances for d_{fea} and those for d_{inp} approach each other as the rate R grows. The upper bound $D_{\mathrm{inp},G}$ reasonably approximates the quantizer performance by the preimages, and it indicates that, in the *two*-dimensional case (Figure 1a), the results for $R = 2$ and 3 bits can be improved by at least about 1 bit.

At low distortion levels, each source output should be reconstructed within a small neighborhood in the feature space where we can find another point y in the input space whose feature map $\phi(y)$ is sufficiently close to the reconstruction. This suggests that the rate-distortion function of feature space reconstruction is well approximated by the rate-distortion function of input space reconstruction. In other words, combining multiple input points to make a reconstruction in feature space does not do any good for reducing distortion and only a single input point is enough when it is mapped into feature space. Hence, the rate-distortion bounds of input space reconstruction may be informative at low distortion levels.

In the 10-dimensional case (Figure 1b), the distortion in the test data set is close to $D_{\mathrm{inp},G}(R)$ or above it at high rates. This may be due to overfitting of the kernel K-means to the training data set of the size, 4000. That is, as the the rate grows, the distortion in the training data set decreases and the discrepancy between the distortions in the training and test sets increases.

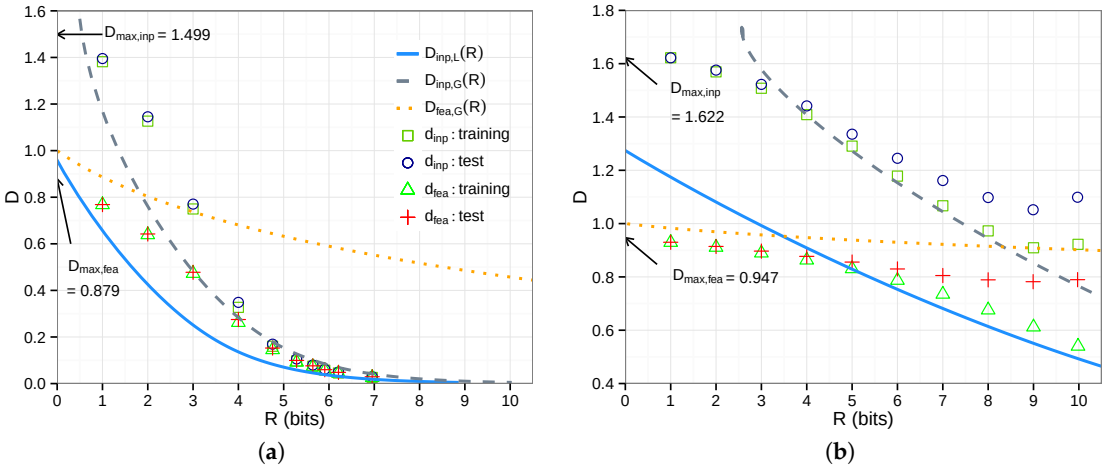

Figure 1. Rate-distortion bounds and quantizer performances for (**a**) $m = 2$ and (**b**) $m = 10$ [17].

To examine the asymptotic behavior of $R_{\mathrm{inp}}(D)$ discussed in Section 4.3, we computed $R_{\mathrm{inp},L}(D)$ and $R_{\mathrm{inp},G}(D)$ for small D, that is, for large s. As well as the Gaussian kernel Equation (29), which has $\alpha = 2$ in Equation (19), we applied the Laplacian kernel,

$$K(x, y) = e^{-\gamma \|x-y\|},$$

which corresponds to $\alpha = 1$. The kernel parameter of the Laplacian kernel was set to the square root of the value used in the Gaussian kernel.

The rate-distortion bounds, $R_{\mathrm{inp},L}(D)$ and $R_{\mathrm{inp},G}(D)$ divided by $-(\log D)/2$ for small distortion levels are shown in Figure 2a,b and for $m = 2$ and $m = 10$, respectively. We can see that, in each case, the ratio tends to $2m/\alpha$, that is, the rate-distortion dimension evaluated in Equation (22) as $D \to 0$. For the distortion levels smaller than those presented in Figure 2, the ratios start oscillating due to the errors of numerical integrations.

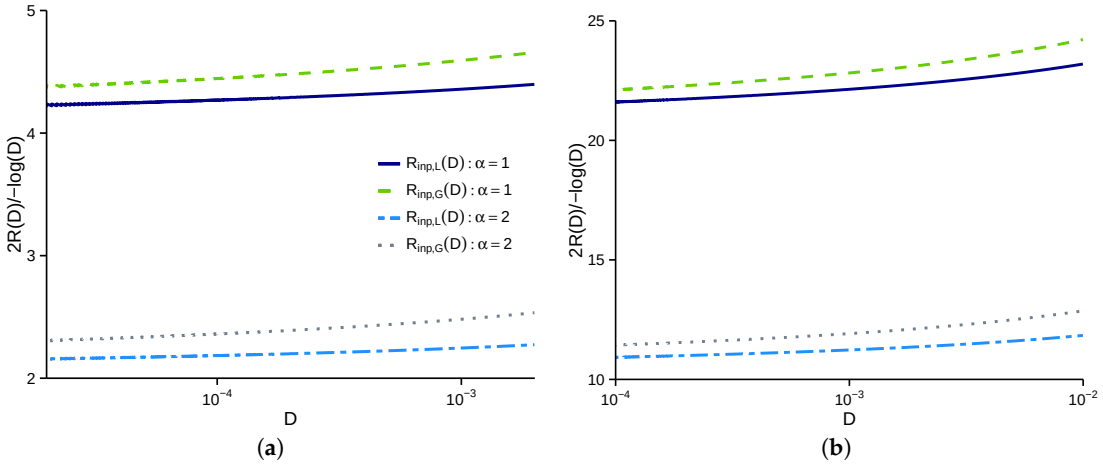

Figure 2. The ratios between the rate-distortion bounds and $-(\log D)/2$ for (**a**) $m = 2$ and (**b**) $m = 10$. The bounds are for the Laplacian kernel ($\alpha = 1$) and the Gaussian kernel ($\alpha = 2$).

5.2. Image Data

We carried out a similar evaluation of the rate-distortion bounds and quantizer performances for a grayscale image data set extracted from the *COIL20* data set [18]. We used the first category from 20 categories of images, which consisted of 72 images of size 32×32. Dividing each 32×32 image into small patches of size 2×2 ($m = 4$), we obtained 256 data from each image, and 18,432 data in total. Removing duplicate data points, we finally obtained 13,368 data. We used first 2048 data as the training data and the remaining 11,320 data as the test data. The training data set was also used for approximating expectations of kernel functions required to compute $R_{\mathrm{fea}}(D)$, and the first $n = 256$ data points were used as the sample data in the definition of d_{fea}. We evaluated only the upper bounds, $R_{\mathrm{fea},G}$ and $R_{\mathrm{inp},G}$, since the lower bound $R_{\mathrm{inp},L}$ requires estimating the source entropy from empirical data, which depends heavily on the estimation method, and hence is to be addressed more in detail.

Each dimension was normalized so that it has mean 0 and variance 1. Hence, v_p in $R_{\mathrm{inp},G}$ was approximated by the empirical variance, 1. The boundary B in $R_{\mathrm{inp},G}$ was approximated by the maximum norm of the training data points.

The upper bounds and quantizer performances are presented in Figure 3. Although the upper bounds are loose and above the respective quantizer performances, the upper bound $D_{\mathrm{inp},G}(R)$ is roughly predictive of the quantizer performance in the input space, and so does $\min\{D_{\mathrm{inp},G}(R), D_{\mathrm{fea},G}(R)\}$ for the reconstruction in the feature space.

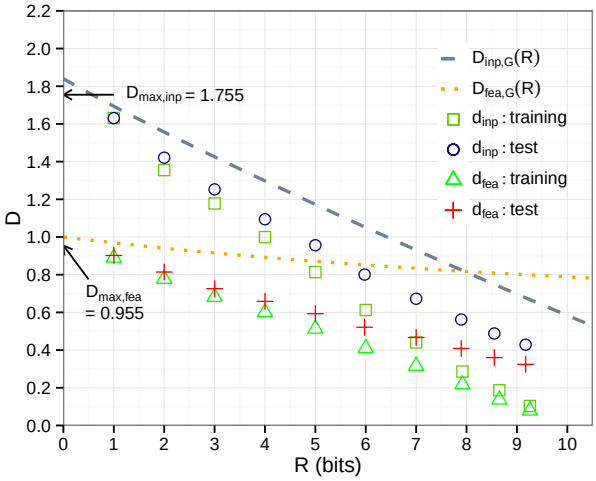

Figure 3. Upper bounds of the rate-distortion functions and quantizer performance for image data.

6. Conclusions

In this paper, we have shown upper and lower bounds for the rate-distortion functions associated with kernel feature mapping. As suggested in Section 5, the upper bound for the reconstruction in feature space is informative at high distortion levels while the bounds for the reconstruction in input space are informative at low distortion levels. We have also evaluated the rate-distortion dimension of sources with bounded support under kernel-based distortion measures, which shows the asymptotic behavior of the rate-distortion function. Our future directions include deriving tighter bounds and exact evaluation of the rate-distortion function in some special cases. In particular, it is an important undertaking to derive a lower bound to the rate-distortion function of the reconstruction in feature space.

Acknowledgments: The author would like to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported in part by the Japan Society for the Promotion of Science (JSPS) grants 25120014, 15K16050, and 16H02825.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proof of Theorem 1

Proof. Let $q^*(y|x)$ be the conditional density for $x \in \mathcal{X}$ that achieves the infimum of $R_{\mathrm{inp}}(D)$. Then, for $Y \sim \int q^*(y|x)p(x)dx$, it holds that $I(X;Y) = R_{\mathrm{inp}}(D)$ and

$$E\left[\|\phi(X) - \phi(Y)\|^2\right] \le D. \tag{A1}$$

Since the input space $\mathcal{X}$ is bounded and separable, and the kernel function K is continuous, for any $\varepsilon > 0$ and $y \in \mathcal{X}$, there exist coefficients $\{\alpha_i(y)\}$ such that

$$\left\|\phi(y) - \sum_{i=1}^{n} \alpha_i(y)\phi(x_i)\right\| \le \frac{\varepsilon}{3\sqrt{D}} \tag{A2}$$

holds when n is sufficiently large.

Let $\boldsymbol{\alpha}(y) = (\alpha_1(y), \ldots, \alpha_n(y))^T$ and

$$q^*(\boldsymbol{\alpha}|x) = \int \delta(\boldsymbol{\alpha} - \boldsymbol{\alpha}(y))q^*(y|x)dy,$$

where δ is Dirac's delta function. Then, for $\boldsymbol{A} \sim \int q^*(\boldsymbol{\alpha}|x)p(x)dx$, it follows from the triangle inequality that

$$\begin{aligned} E\left[d_{\mathrm{fea}}(X,\boldsymbol{A})\right] &= E\left[\left\|\phi(X)-\sum_{i=1}^{n}\alpha_i(Y)\phi(x_i)\right\|^2\right] \\ &\leq E\left[\left\|\phi(X)-\phi(Y)\right\|^2\right]+2E\left[\left\|\phi(X)-\phi(Y)\right\|\left\|\phi(Y)-\sum_{i=1}^{n}\alpha_i(Y)\phi(x_i)\right\|\right] \\ &\quad +E\left[\left\|\phi(Y)-\sum_{i=1}^{n}\alpha_i(Y)\phi(x_i)\right\|^2\right], \end{aligned}$$

and hence

$$\begin{aligned} E\left[d_{\mathrm{fea}}(X,\boldsymbol{A})\right] &\leq D+\frac{2\varepsilon}{3}+\frac{\varepsilon^2}{9D} && \text{(A3)} \\ &\leq D+\varepsilon. && \text{(A4)} \end{aligned}$$

To obtain Inequality (A3), we used Equations (A1) and (A2), and Jensen's inequality,

$$\begin{aligned} E\left[\sqrt{\left\|\phi(X)-\phi(Y)\right\|^2}\right] &\leq \sqrt{E\left[\left\|\phi(X)-\phi(Y)\right\|^2\right]} \\ &\leq \sqrt{D}. \end{aligned}$$

Thus, from Equation (A4) and the data-processing inequality [7], we have

$$R_{\mathrm{fea}}(D+\varepsilon)\leq I(X;\boldsymbol{A})\leq I(X;Y)=R_{\mathrm{inp}}(D),$$

which completes the proof. □

Appendix B. Proof of Theorem 5

Proof. The mean and covariance matrix of the random vector $\boldsymbol{A}\sim M_p(\boldsymbol{\alpha})$ are

$$\begin{aligned} E[\boldsymbol{A}] &= \tilde{\boldsymbol{K}}^{-1}\int \boldsymbol{k}(x)p(x)dx \\ Cov[\boldsymbol{A}] &= E\left[\boldsymbol{A}\boldsymbol{A}^T\right]-E\left[\boldsymbol{A}\right]E\left[\boldsymbol{A}\right]^T \\ &= \left\{\frac{1}{2s}\boldsymbol{I}+\tilde{\boldsymbol{K}}^{-1}\int \boldsymbol{k}(x)\boldsymbol{k}(x)^Tp(x)dx\right\}\tilde{\boldsymbol{K}}^{-1}-\tilde{\boldsymbol{K}}^{-1}\int \boldsymbol{k}(x)p(x)dx\int \boldsymbol{k}(x)^Tp(x)dx\tilde{\boldsymbol{K}}^{-1} \\ &= \left\{\frac{1}{2s}\boldsymbol{I}+\tilde{\boldsymbol{K}}^{-1}\boldsymbol{C}\right\}\tilde{\boldsymbol{K}}^{-1}, \end{aligned}$$

where $\boldsymbol{C}$ is defined by Equation (28).

Thus, the maximum entropy principle of the Gaussian distribution implies that the differential entropy $h(M_p)$ is upper-bounded by

$$h(M_p)\leq \frac{n}{2}\log\left[(2\pi e)\left|\left\{\frac{1}{2s}\boldsymbol{I}+\tilde{\boldsymbol{K}}^{-1}\boldsymbol{C}\right\}\tilde{\boldsymbol{K}}^{-1}\right|^{\frac{1}{n}}\right].$$

Combining this inequality with Equations (24) and (26), we have

$$R_{\mathrm{fea}}(D_s)\leq \frac{1}{2}\log\left|\boldsymbol{I}+2s\tilde{\boldsymbol{K}}^{-1}\boldsymbol{C}\right|.$$

Solving Equation (25) with respect to $2s$ and substituting it into the above expression, we obtain the upper bound in Equation (27). □

References

1. Schölkopf, B.; Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2001.

2. Aizerman, M.A.; Braverman, E.A.; Rozonoer, L. Theoretical foundations of the potential function method in pattern recognition learning. *Autom. Remote Control* **1964**, *25*, 821–837.
3. Berger, T. *Rate Distortion Theory: A Mathematical Basis for Data Compression*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1971.
4. Girolami, M. Mercer kernel-based clustering in feature space. *IEEE Trans. Neural Netw.* **2002**, *13*, 780–784.
5. Filippone, M.; Camastra, F.; Masulli, F.; Rovetta, S. A survey of kernel and spectral methods for clustering. *Pattern Recognit.* **2008**, *41*, 176–190.
6. Schölkopf, B.; Mika, S.; Burges, C.J.C.; Knirsch, P.; Müller, K.R.; Ratsch, G.; Smola, A.J. Input space versus feature space in kernel-based methods. *IEEE Trans. Neural Netw.* **1999**, *10*, 1000–1017.
7. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley Interscience: Hoboken, NJ, USA, 1991.
8. Gray, R.M. *Entropy and Information Theory*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2011.
9. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer: Berlin/Heidelberg, Germany, 2008.
10. Linkov, Y.N. Evaluation of ϵ-entropy of random variables for small ϵ. *Probl. Inf. Transm.* **1965**, *1*, 18–26.
11. Linder, T.; Zamir, R. On the asymptotic tightness of the Shannon lower bound. *IEEE Trans. Inf. Theory* **1994**, *40*, 2026–2031.
12. Koch, T. The Shannon lower bound is asymptotically tight. *IEEE Trans. Inf. Theory* **2016**, *62*, 6155–6161.
13. Kawabata, T.; Dembo, A. The rate-distortion dimension of sets and measures. *IEEE Trans. Inf. Theory* **1994**, *40*, 1564–1572.
14. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*; The MIT Press: Cambridge, MA, USA, 2005.
15. Fukunaga, K.; Hostetler, L. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory* **1975**, *21*, 32–40.
16. Comaniciu, D.; Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 603–619.
17. Watanabe, K. Rate-distortion analysis for kernel-based distortion measures. In Proceedings of the Eighth Workshop on Information Theoretic Methods in Science and Engineering, Copenhagen, Denmark, 24–26 June 2015.
18. Nene, S.A.; Nayar, S.K.; Murase, H. Columbia Object Image Library (COIL-20). Available online: http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php (accessed on 4 July 2017).