CHAPTER 2
Renyi's Entropy, Divergence and Their Nonparametric Estimators

Chapter Coauthors: Dongxin Xu, and Deniz Erdogmus

2.1. Introduction

It is evident from Chapter 1 that Shannon's entropy occupies a central role in information-theoretic studies. Yet, the concept of information is so rich that perhaps there is no single definition that will be able to quantify information properly. Moreover, from an engineering perspective, one must estimate entropy from data which is a nontrivial matter. In this book we concentrate on Alfred Renyi's seminal work on information theory to derive a set of estimators to apply entropy and divergence as cost functions in adaptation and learning. Therefore, we are mainly interested in computationally simple, nonparametric estimators that are continuous and differentiable in terms of the samples to yield well-behaved gradient algorithms that can optimize adaptive system parameters. There are many factors that affect the determination of the optimum of the performance surface, such as gradient noise, learning rates, and misadjustment, therefore in these types of applications the entropy estimator's bias and variance are not as critical as, for instance, in coding or rate distortion theories. Moreover in adaptation one is only interested in the extremum (maximum or minimum) of the cost, with creates independence from its actual values, because only relative assessments are necessary. Following our nonparametric goals, what matters most in learning is to develop cost functions or divergence measures that can be derived directly from data without further assumptions to capture as much structure as possible within the data's probability density function (PDF).

The chapter starts with a review of Renyi's entropy origins, its properties, and interpretations. Then a new estimator for Renyi's quadratic entropy is developed using kernel density estimation. With cost functions for adaptation in mind, the properties of this estimator which is called the *Information Potential* (IP) *estimator* are carefully presented, including its bias and variance. A physical interpretation of the IP is presented which will motivate new adaptation algorithms in Chapter 3.

A brief review of Renyi's divergence and mutual information is also presented, and two divergence measures in probability spaces are introduced that have the great appeal of being computed from combinations of the IP estimator; that is, IP can be readily extended to estimate divergences. A detailed discussion of the algorithms and interpretations of these divergence measures is presented to allow their use in practical applications. This includes two classes of algorithms that speed up the computations to $O(N)$. Furthermore, Appendix A presents a review of entropy estimators along with a review of how the IP can be used in practical problems.

2.2. Definition and Interpretation of Renyi's Entropy

The parametric family of entropies was introduced by Alfred Renyi in the mid 1950s as a mathematical generalization of Shannon entropy [263]. Renyi wanted to find the most general class of information measure that preserved the additivity of statistically independent systems and were compatible with Kolmogorov's probability axioms.

Let us assume a discrete probability distribution P ={p_1, p_2, …, p_N} fulfilling the conditions of $\sum_k p_k = 1, \quad p_k \geq 0$. If one observes the outcome of two independent events with probabilities p and q, additivity of information for independent events requires that the corresponding information $I(\cdot)$ obey Cauchy's functional equation (i.e. the information of the joint event is the sum of the information of each event)

$$I(P \cdot Q) = I(P) + I(Q)\,. \tag{2.1}$$

Therefore, the amount of information produced by knowing that an event with probability p took place could be, apart from a multiplicative factor (normalized by setting $I(1/2) = 1$)

$$I(P) = -\log_2 p\,, \tag{2.2}$$

which is similar to Hartley's amount of information. Let us further assume that the outcomes of some experimental discrete random variable occur with probabilities $p_1, \ldots,$ p_N, and if the kth outcome delivers I_k bits of information then the total amount of information for the set Γ={I_1,….I_N} is

$$I(P) = \sum_{k=1}^{N} p_k I_k \tag{2.3}$$

which can be recognized as Shannon's entropy $H(X)$. However, we have assumed the linear averaging operator in this formulation. In the general theory of means for any monotonic function $g(x)$ with an inverse $g^{-1}(x)$ one can define the general mean associated with $g(x)$ for a set of real values $\{x_k \mid k = 1,\ldots, N\}$ with probabilities of $\{p_k\}$ as

$$g^{-1}\Big(\sum_{k=1}^{N} p_k g(x_k)\Big).$$

Applying this definition to the information $I(P)$, we obtain

$$I(P) = g^{-1}\Big(\sum_{k=1}^{N} p_k g(I_k)\Big)\,, \tag{2.4}$$

where $g(x)$ is a Kolmogorov–Nagumo invertible function [229]. This $g(x)$ is the so called quasi-linear mean and it constitutes the most general mean compatible with Kolmogorov's axiomatics [184]. Renyi then proved that when the postulate of additivity for independent events is applied to Eq. (2.4) it dramatically restricts the class of possible $g(x)$. In fact, only two classes are possible; $g(x) = cx$ with c a constant, which states that for linear $g(x)$ the quasi-linear mean reduces to the ordinary mean and yields the Shannon information measure Eq.(2.3). Hence, Shannon's information is the averaged information in the usual sense, and becomes the simplest of the information measures. The other functional class is $g(x) = c\, 2^{(1-\alpha)x}$ which implies

$$I_\alpha(P) = \frac{1}{1-\alpha}\log\Big(\sum_{k=1}^{N} p_k^\alpha\Big)$$

with $\alpha \neq 1$and $\alpha \geq 0$, and it is called Renyi's information measure of order α, or Renyi's α entropy, denoted as $H_\alpha(X)$. We adopt the term "entropy" since Renyi showed that it also represents the disclosed information (or removed ignorance) after analyzing the expression in a close analogy with Shannon's theory.

At a first glance, the main difference between Shannon and Renyi's entropies is the placement of the logarithm in the expression. In Shannon entropy (Eq. (1.4), the probability mass function (PMF) weights the log(p_k) term, whereas in Renyi's entropy the log is outside a term that involves the α power of the PMF. In order to compare further with Shannon's entropy let us rewrite Renyi's entropy as

$$H_\alpha(X)=\frac{1}{1-\alpha}\log\left(\sum_{k=1}^{N}p_k^\alpha\right)=-\log\left(\sum_{k=1}^{N}p_k^\alpha\right)^{\frac{1}{\alpha-1}}=-\log\left(\sum_{k=1}^{N}p_k p_k^{\alpha-1}\right)^{\frac{1}{\alpha-1}}. \qquad (2.5)$$

We see in Eq. (2.5) that the PMF p_k also weights a term that now is the (α−1) power of the probability mass function. Let us denote the argument of the log as $V_\alpha(X)=\sum_k p_k^\alpha=E[p_k^{\alpha-1}]$ which is called in this book the α *information potential* (IP$_\alpha$) and allows rewriting Eq. (2.5) as

$$H_\alpha(X)=\frac{1}{1-\alpha}\log(V_\alpha(X))=-\log(\sqrt[\alpha-1]{V_\alpha(X)}). \qquad (2.6)$$

At a deeper level, Renyi's entropy measure is much more flexible due to the parameter α, enabling several measurements of uncertainty (or dissimilarity) within a given distribution [177]. Considered as a function of α, $H_\alpha(X)$ is normally called the spectrum of Renyi information and its graphical plot is useful in statistical inference [308]. The value at α = 1 is particularly important because it provides the expected value of the negative log-likelihood (E[−log $p(x)$]) while its derivative with respect to α is proportional to the variance of the log-likelihood function ($\dot{H}_1(X)=-1/2\,\mathrm{var}(\log p(x)))$. Due to this fact it is possible to derive an index of the intrinsic shape of the PDF as $S(X)=-2\dot{H}_1(X)$ which has more statistical power than kurtosis and can be used as a partial order for the tails of distributions.

To find the most fundamental (and possibly irreducible) set of properties characterizing Renyi's information it is desirable to axiomatize it. Various axiomatizations have been proposed [265, 1]. For our purpose the most convenient set of axioms is the following [340].

1. The entropy measure $H(p_1,....,p_N)$ is a continuous function of all the probabilities p_k, which means that a small change in probability distribution will only result in a small change in the entropy.

2. $H(p_1,...,p_N)$ is permutationally symmetric; that is, the position change of any two or more p_k in $H(p_1,....,p_N)$ will not change the entropy value. Actually, the permutation of any p_k in the distribution will not change the uncertainty or disorder of the distribution and thus should not affect the entropy.

3. $H(1/n,....,1/n)$ is a monotonic increasing function of n. For an equiprobable distribution, when the number of choices increases, the uncertainty or disorder increases, and so does the entropy measure.

4. Recursivity: If an entropy measure satisfies

$$H(p_1, p_2, \ldots, p_N) = H(p_1 + p_2, p_3, \ldots, p_N) + (p_1 + p_2)^{\alpha} H(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2})$$

then it has a recursivity property. It means that the entropy of N outcomes can be expressed in terms of the entropy of $N - 1$ outcomes plus the weighted entropy of the combined two outcomes.

5. Additivity: If $p = (p_1, \ldots, p_N)$ and $q = (q_1, \ldots, q_N)$ are two independent probability distributions, and the joint probability distribution is denoted by $p \cdot q$, then the property $H(p \cdot q) = H(p) + H(q)$ is called additivity.

The table compares Renyi's entropy property versus Shannon for these axioms.

Properties	(1)	(2)	(3)	(4)	(5)
Shannon	yes	yes	yes	yes	yes
Renyi	yes	yes	yes	no	yes

Notice that Renyi's recursivity property differs from Shannon's recursivity, so we entered *no* in Property (4) to make this fact clear. Further properties of Renyi's entropy were studied extensively in [265, 1]. We list here a few of the key ones.

(a) $H_\alpha(X)$ is nonnegative: $H_\alpha(X) \geq 0$.

(b) $H_\alpha(X)$ is decisive: $H_\alpha(0, 1) = H_\alpha(1, 0) = 0$.

For $\alpha \leq 1$ Renyi's entropy is concave. For $\alpha > 1$ Renyi's entropy is neither pure convex nor pure concave. It loses concavity for $\alpha > \alpha^* > 1$ where α^* depends on N as $\alpha^* \leq 1 + \ln(4)/\ln(N-1)$.

(d) Because

$$\frac{\alpha - 1}{\alpha} H_\alpha(X) \leq \frac{\beta - 1}{\beta} H_\beta(X), \quad \alpha \geq \beta,$$

$(\alpha - 1)H_\alpha(X)$ is a concave function of p_k.

(e) $H_\alpha(X)$ is a bounded, continuous and nonincreasing function of α.

(f) For $\alpha \in R$; $H_\alpha(A \cap B) = H_\alpha(A) - H_\alpha(B|A)$ with $H_\alpha(B \mid A) = g^{-1}(\sum_k \rho_k(\alpha) g(H_\alpha(B \mid A = A_k)))$,

which can be interpreted as the conditional entropy with $\rho_k(\alpha) = p_k^\alpha / \sum_k p_k^\alpha$ and g an invertible and positive function in [0,1).

(g) $H_z(X)$ with $z = \alpha + j\omega$ is analytic in the entire complex plane except the negative real axis. Therefore the singularity at $\alpha = 1$ is not essential and we obtain $\lim_{\alpha \to 1} H_\alpha(X) = H_S(X)$.

The ambiguous concavity property of Renyi's entropy in (b) makes it incompatible with the requirements of physical entropy (unlike Shannon's) when expressed as a function of the pertinent p_k. The implications of (g) are far reaching. It can be shown that if we perform an analytical continuation of α in the complex domain (e.g., $z = \alpha + j\omega$), $H_z(X) = \sum_{k=1}^{N} p_k^z$ is analytic except in the negative real axis. More specifically, if we make $z = 1 + re^{j\omega}$, $H_z(X)$ is analytic in the interior of the circle of radius r so it is also analytic at $z = \alpha = 1$. Therefore Renyi's entropy is differentiable at $z = 1$ to all orders. With this proof, Shannon entropy can be uniquely determined from the behavior of (analytically continued) Renyi's entropy in the vicinity of $z = 1$. Therefore from a strict mathematical point of view, Shannon entropy is not a special information measure deserving separate axiomatization but a member of Renyi's wide class of entropies embraced by a single unifying axiomatic [168]. Despite its formal origin Renyi's entropy proved important in a variety of practical applications: coding theory [44], statistical inference [236], quantum mechanics (as an estimator for von Neumann entropy) [32], chaotic dynamical systems [120], multifractal analysis [168], and as a measure of diversity in economics [135].

Geometric Interpretation of Renyi's Entropy

Before we actually start the derivation of the estimators, we investigate further the role of α by providing a geometric picture of Renyi's entropy that is very useful to describe this family of entropies. Probability mass functions can be visualized geometrically as points in a vector space with the axis given by the random variables called the *simplex*. The simplex Δ_N consists of all possible probability distributions for an N multidimensional random variable; that is,

$$\Delta_N = \{ p = (p_{1,} .., p_N)^T \in R^N ,\ \ p_i \geq 0,\ \sum_i p_i = 1,\ \forall i \}$$

For instance, for three variables (x, y, z), the space of all such distributions is an equilateral triangle with vertices at (1,0,0), (0,1,0), (0,0,1) (a convex subset of R^3). Figure 2.1 shows the simplex for $N = 2$ and $N = 3$.

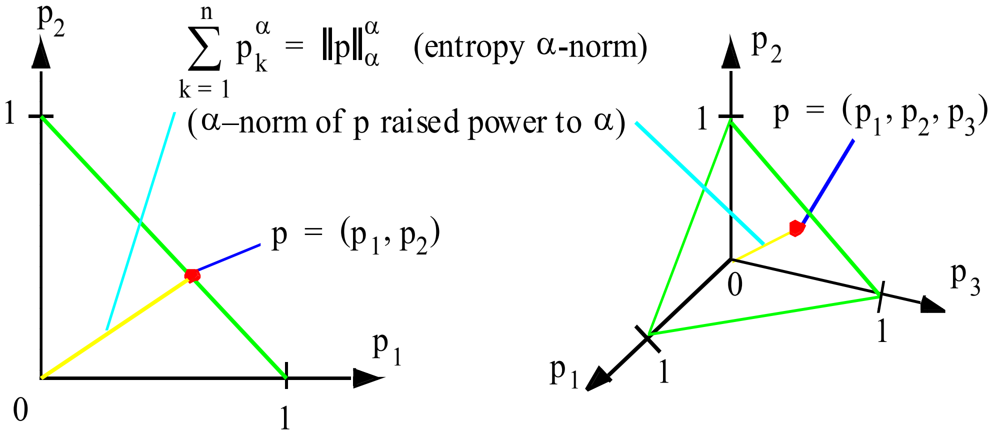

p_2
$\sum_{k=1}^{n} p_k^{\alpha} = \|p\|_{\alpha}^{\alpha}$ (entropy α-norm)
(α–norm of p raised power to α)
1
$p = (p_1, p_2)$
p_1
0
1
p_2
1
$p = (p_1, p_2, p_3)$
0
p_3
1
p_1
1

Figure 2.1. The simplex for N = 2, 3 and the entropy α norm.

Any point in the simplex is a different PMF and has a different distance to the origin. If one defines the PMF α-norm as

$$\|p(x)\|_{\alpha} = \sqrt[\alpha]{\sum_{k=1}^{N} p_k^{\alpha}} = \sqrt[\alpha]{V_{\alpha}(X)} ,$$

that is, the α-information potential $V_\alpha(x)$ can be interpreted as the α power of the PMF α-norm. Specifically, Renyi's α entropy takes the $\alpha - 1$ root of $V_\alpha(x)$ and rescales it by the negative of the logarithm as specified in Eq. (2.6). Therefore α specifies in the simplex the norm to measure the distance of $p(x)$ to the origin. As is well known from the theory of norms [40], the free parameter α specifying the norm changes the importance of small values versus large values in the set. Three α cases are of special interest: H_0 is the logarithm of the number of nonzero components of the distribution and is known as Hartley's entropy. H_∞ can be thought of as $\lim_{\alpha\to\infty} H_{R^\alpha} = H_\infty$ with $H_\infty = -\log(\max_k(p_k))$ which is called the Chebyshev entropy [177]. The most interesting special case is obtained for $\lim_{\alpha\to 1} H_\alpha = H_S$ which means that Shannon's entropy is the limiting case of a 1- norm of the probability mass function $p(x)$. Actually, the 1-norm of any probability density is always 1 by definition, which will give 0/0 in Eq (2.6). Using the l'Hôpital rule we can proceed and evaluate Eq. (2.6) as

$$\lim_{\alpha\to 1} H_\alpha(X) = \lim_{\alpha\to 1} \frac{\frac{d}{d\alpha}\log\sum_{k=1}^{N} p_k^\alpha}{\frac{d}{d\alpha}(1-\alpha)} = \frac{\left(\sum_{k=1}^{N}\log p_k \cdot p_k^\alpha\right)\left(\sum_{k=1}^{N} p_k^\alpha\right)^{-1}\Bigg|_{\alpha=1}}{-1} = H_S(X) \qquad (2.7)$$

so, in the limit, Shannon's entropy can be regarded as the functional value of the 1-norm of the probability density.

Renyi's entropy is a scalar that characterizes densities, thus it is also interesting to display the contours of equal Renyi's entropy in the simplex (Figure 2.2) for several α.

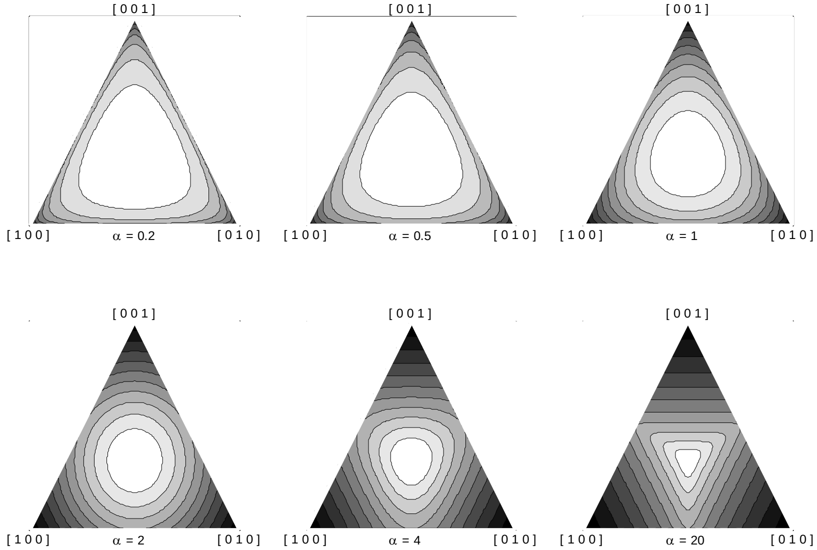


Figure 2.2. Isoentropy contour in the N = 3 probability simplex for different α values.

In order to illustrate how Renyi's entropy evaluation behaves in the simplex, we plot the isoentropy contours as a function of α. Notice that for α close to zero the values inside the simplex change very little, and the Shannon case basically preserves the shape of these contours except that there is a more visible change. Observe that for α = 2 the contours are circular, meaning a l_2-norm to the origin. For higher values of α the contours rotate by 180 degrees and emphasize changes with respect to the central point when α increases.

When $\alpha > 1$, Renyi's entropy H_α are monotonic decreasing functions of IP_α. So, in this case, the entropy maximization is equivalent to IP minimization, and the entropy minimization is equivalent to IP maximization.

When $\alpha \leq 1$, Renyi's entropy H_α are monotonic increasing functions of the V_α. So, in this case, the entropy maximization is equivalent to IP maximization, and the entropy minimization is equivalent to IP minimization.

Renyi's Quadratic Entropy H_2

H_2 is of particular interest in this book and it is a monotonic decreasing function of the $\alpha = 2$ information potential V_2 (V for short) of the PMF $p(x)$. H_2 implicitely uses an Euclidean distance from the point $p(x)$ in the simplex to the origin.

$$H_2(X) = -\log(\sum_k p_k^2)\,. \tag{2.8}$$

In the particle physics literature, the second moment of the probability mass function is known as the index of coincidence or *purity* (because it vanishes if the state of the particle is pure) [32]. The *linear entropy* is defined as $H_L(X) = 1 - p^2(x)$ (which is in fact the Havrda - Charvat [138] or Tsallis entropy of second order [320]), but in Renyi's case, the logarithm is used instead. In econometrics, Renyi's quadratic entropy has been used to quantify diversity [135]. Because H_2 is a lower bound of Shannon's entropy, it might be more efficient than Shannon's entropy for entropy maximization.

One aspect that we would like to stress after the presentation of the geometric picture of Renyi's entropy is the fact that the argument of the log in Renyi's quadratic entropy, $V_2 = E[p(x)]$ has meaning in itself as the expected value of the PMF. Equivalently, if one considers the PMF a nonlinear function of the random variable x and defines the transformed random variable $\xi = p(x)$, IP_2 (IP for short) becomes the expected value of ξ. The argument of the log in $H_2(x)$ is central to our studies. In fact, we show that in optimization (parameter adaptation) the logarithm is irrelevant (because it is a monotonic function and therefore does not affect the location of the extremum of the cost function in the space of the system parameters) and is dropped almost from the beginning of our adaptive system studies. This is unthinkable in communication theory, because there the fundamental issue is the additivity of information, which is intrinsically linked to the logarithmic function.

Some authors [32] define $f_\alpha(\mathbf{p}) = \sum_{k=1}^{N} p_k^\alpha$ as the *α moment of the probability mass function,* which is Schur concave for α < 1 and Schur convex for α > 1. Therefore Renyi's entropy is a function of the moment of the vector variable $\mathbf{p} = [p_1, p_2,..., p_N]$. Moreover, the moments $f_\alpha(p_1, p_2,\ldots, p_N)$ for $\alpha = 2, \ldots, N$ define the vector $\mathbf{p}$ up to a permutation of its components, which means that the spectrum of Renyi's entropies defines the probability mass function in a similar manner as the characteristic function expansion. The α moments also relate Renyi's to von Neumann's entropy [326]. It is important not to confuse the moments of the PMF with the moments of the data, therefore we prefer to use the information potential terminology, which also has a powerful analogy as we discuss later. Figure 2.3 shows the relation between the 2-norm of the PMF (mean of ξ) and the mean of the data, which should be obvious.

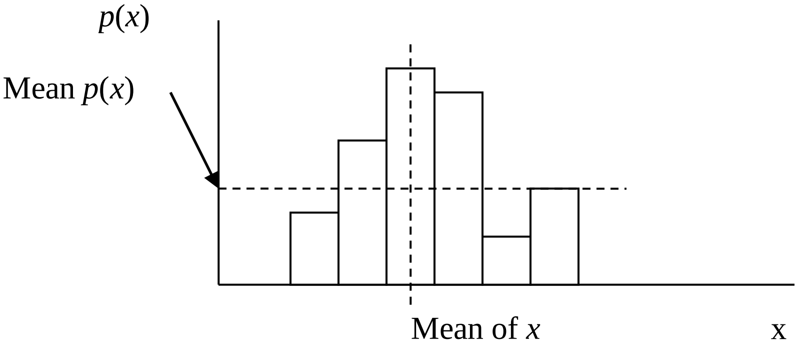


Figure 2.3. Relation between mean of *x* and mean of *p*(*x*).

Renyi's Entropy of Continuous Variables

Renyi's entropy can also be defined for continuous random variables. Let *p*(*x*) be the continuous PDF defined in [0,1]. The integrated probability is

$$p_{n,k} = \int_{k/n}^{(k+1)/n} p(x)dx, \quad k = 0,1,..., n-1$$

and by defining the discrete mass function $P_n=\{p_{n,k}\}$ it is possible to show [265] that

$$H_\alpha(X) = \lim_{n\to\infty}(I_\alpha(P_n) - \log n) = \frac{1}{1-\alpha}\log\int p^\alpha(x)dx. \qquad (2.9)$$

This is very similar to the Shannon case, showing that the differential Renyi's entropy can be negative for $\alpha \le 1$. Indeed log(*n*) can be thought as the entropy of the uniform distribution, and so the continuous entropy is the gain obtained by substituting the uniform distribution by the experimental samples P_n. The generalization for multidimensions proceeds along the same arguments, preserving the functional form of Eq. (2.9). Quadratic Renyi's entropy for continuous random variables reads

$$H_2(X) = -\log\int p^2(x)dx. \qquad (2.10)$$

We use capital *H* for differential entropy throughout this book.

2.3. Quadratic Renyi's Entropy Estimator

As already stated in Chapter 1, in experimental science, one is faced with the issue of estimating entropy directly from samples in a nonparametric way because it is often not prudent to advance with a parametric PDF model. In such cases we have to resort to a nonparametric estimation. But instead of first estimating the PDF and then computing its entropy, here we seek the direct approach of estimating quadratic Renyi's entropy from samples by estimating *E*[*p*(*X*)], which is a scalar. In adaptive systems we are mostly interested in continuous random variables, and this is the case on which we concentrate from this point on.

Recall the definition of quadratic entropy given in Eq. (2.10) for the continuous random variable *X*. Suppose we have *N* independent and identically distributed (i.i.d.) samples $\{x_1, \ldots, x_N\}$ from this random variable. The kernel (Parzen) estimate of the PDF [241] using an arbitrary kernel function $\kappa_\sigma(.)$ is given by

$$\hat{p}_X(x) = \frac{1}{N\sigma}\sum_{i=1}^{N}\kappa(\frac{x - x_i}{\sigma}) \tag{2.11}$$

where σ is the kernel size or bandwidth parameter. This kernel function has to obey the following properties [300].

1. $\kappa(x) \geq 0.$
2. $\int_R \kappa(x)dx = 1.$
3. $\lim_{x\to\infty} |x\kappa(x)| = 0.$

Normally one uses a symmetric normalized kernel that peaks at the sample and for our purposes it must be continuous and differentiable (reasons are discussed later). Kernel density estimation is a–well-studied topic and the use of kernels has been widespread since the seminal work of Rosenblatt and Parzen. The quality of estimators is normally quantified by their bias and variance, and for kernel estimation they are respectively given by [241] (^ denotes estimated quantities)

$$\begin{aligned} &Bias(\hat{p}_\sigma(x)) = E[\hat{p}_\sigma(x)] - p(x) \approx \sigma^2/2p''(x)\mu(K) \\ &Var(\hat{p}_\sigma(x)) = E[(\hat{p}_\sigma(x) - E[p_\sigma(x)])^2] \approx \frac{1}{N\sigma}\|K\|_2^2\, p(x), \qquad N\sigma \to \infty \end{aligned} \tag{2.12}$$

where μ(*K*) and $\|K\|^2$ are constants given by the specific kernel utilized, and p'' is the second derivative of the PDF. As one can see in Eq. (2.12) the kernel size affects the bias and the variance in opposite ways, so the best kernel size is a compromise between bias and variance of the estimator. It is well known from Parzen's seminal work [241] that the class of kernel estimators is asymptotically unbiased when the kernel size tends to zero (i.e., the kernel approaches a Dirac delta function), and consistent in quadratic mean when the number of samples increases to infinite (the product $N\sigma$ must tend to infinite). Moreover, one can show that the mean square error between the true and estimated PDF can decrease for the optimal kernel size at a rate as high as $N^{-4/5}$ for scalar variables, which is close to the best possible (1/*N*). For symmetric kernels such as the Gaussian it is typically $N^{-2/5}$.

The difficulty of density estimation, in particular in high dimensions and with few samples, is that one wants to obtain a reasonable estimate for all the points in the domain. This is an ill-posed problem (see Appendix A). However, for $V_2(X)$, we are only interested in estimating a single number $E[p(x)]$. Assuming Gaussian kernels Eq. (1.50), $G_\sigma(.)$, with standard deviation σ and substituting this in the quadratic entropy expression Eq. (2.10), we get after straightforward substitutions the estimator

$$\hat{H}_2(X) = -\log \int_{-\infty}^{\infty} \left(\frac{1}{N} \sum_{i=1}^{N} G_\sigma(x - x_i) \right)^2 dx = -\log \frac{1}{N^2} \int_{-\infty}^{\infty} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} G_\sigma(x - x_j) \cdot G_\sigma(x - x_i) \right) dx$$

$$= -\log \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{-\infty}^{\infty} G_\sigma(x - x_j) \cdot G_\sigma(x - x_i) dx = -\log(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i)). \qquad (2.13)$$

The result is easily obtained by noticing that the integral of the product of two Gaussians is *exactly evaluated* as the value of the Gaussian computed at the difference of the arguments and whose variance is the sum of the variances of the two original Gaussian functions. Other kernel functions, however, do not result in such convenient evaluation of the integral because the Gaussian maintains the functional form under convolution. Nevertheless, any positive definite function that peaks at the origin (most kernels) might still be used in the estimation, but the expressions become a bit more complicated. We named the argument of the log in Eq. (2.13) (i.e., the kernel estimator of the 2-norm of the PMF (or PDF)) the *quadratic information potential estimator* (simply IP when there is no confusion) for reasons that become apparent later.

Information Potential for Entropy Estimation

The argument of the logarithm in quadratic Renyi's entropy that has been called the information potential can be estimated directly from data as

$$\hat{H}_2(X) = -\log(\hat{V}_2(X)) \qquad \hat{V}_{2,\sigma}(X) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i) \qquad (2.14)$$

where $\hat{V}_{2,\sigma}(X)$ is the quadratic IP estimator that depends on σ. Let us compare this with the conventional way that entropy is estimated from data. In practical cases, the estimation of Shannon or Renyi's entropy directly from data will follow the route:

$$data \rightarrow pdf \quad estimation \rightarrow integral \quad estimation\,.$$

Notice that entropy is a scalar, but as an intermediate step one has to estimate a function (the PDF), which is much harder in high dimensional spaces. With quadratic Renyi's entropy and the IP (i.e. *V*(*X*)) we bypassed the explicit need to estimate the PDF; that is, the calculations follow the path

$$data \rightarrow IP(\hat{V}_2(X)) \rightarrow algebra\,.$$

Eq. (2.14) is one of the central results of information-theoretic learning because it shows that the Information Potential, which is a scalar, can be estimated directly from samples with an exact evaluation of the integral over the random variable for Gaussian kernels. Eq. (2.14) shows that the IP is only a function of sample pairs, instead of the PDF shape. This is similar to the conventional estimators of the mean and the variance that work directly with the samples irrespective of the PDF, but unfortunately here the estimator has a free parameter and it shares the properties of kernel density estimation.

There are two important implications of Eq. (2.14). As is apparent, the variance of the Gaussian (also called the kernel size or bandwidth) is a free parameter that needs to be selected by the user. Therefore, when the IP is estimated, the resulting values of entropy depend on the kernel size selected, which is also a crucial problem in density estimation [300]. The estimated values of the IP have little absolute meaning due to this kernel size dependence, but it gauges performance in a relative sense when comparing data generated with the same set of parameters. In learning (the main purpose of this book) the system parameters depend only on the cost function's extremum location in parameter space, not of the cost's actual value, so the IP dependence on kernel size is more manageable than for applications that require the actual value of the estimated quantity.

The way we interpret the kernel bandwidth is as a *scale parameter* for the analysis. It has to be selected according to the data dynamic range and number of samples to make the estimation of the entropy meaningful. Silverman's rule [300] is

$$\sigma_{opt} = \sigma_X \left(4N^{-1}(2d+1)^{-1}\right)^{\frac{1}{(d+4)}}, \tag{2.15}$$

where N is the number of samples, d is the data dimensionality, and σ_X is the data standard deviation. Although Eq. (2.15) was derived for Gaussian distributions it is sufficient for most of our applications. The bandwidth selection is treated more thoroughly in Appendix A. To summarize, we want to say that the existence of this free parameter is a double-sided sword: it provides flexibility in the application of the methodology to real data; but on the other hand it either requires a selection criterion or a scanning over σ because the effect of the kernel size is much harder to quantify than the scale in wavelet decompositions or frequency in Fourier analysis that also contain a free parameter in their definitions. More generally, it shows the functional nature of entropy estimation using kernels.

The second implication is that the estimator is $O(N^2)$, which may create computation bottlenecks for large datasets. This is the price we have to pay to estimate entropy with the IP when compared with mean and variance. Indeed, both the mean and the variance estimators work with a single sample at a time (in fact the variance also requires pairwise computation but one of the elements of the pair is the mean that can be computed a priori), but if we are interested in qualifying the "shape" of the PDF with Renyi's second-order entropy, pairwise interactions are necessary. We show later in this chapter how the fast Gauss transform and the incomplete Cholesky decomposition solve this problem with algorithms that are $O(N)$.

Extended Estimator for α-Renyi's Entropy

It turns out that the pairwise interaction model can be generalized from an estimator of Renyi's quadratic entropy to all $\alpha \neq 1$. In essence $\hat{H}_2$ is the centerpiece for nonparametric kernel estimation of Renyi's entropy as we show below. Consider the definition of Renyi's order-α entropy in Eq. (2.9), which can also be written with an expectation operator as

$$H_\alpha(X) \overset{\Delta}{=} \frac{1}{1-\alpha}\log\int_{-\infty}^{\infty} p_X^\alpha(x)dx = \frac{1}{1-\alpha}\log E_X\left[p_X^{\alpha-1}(X)\right]. \tag{2.16}$$

Approximating the expectation operator with the sample mean as is commonly done in density estimation [300], we get

$$H_\alpha(X) \approx \hat{H}_\alpha(X) = \frac{1}{1-\alpha}\log\frac{1}{N}\sum_{j=1}^{N} p_X^{\alpha-1}(x_j). \tag{2.17}$$

Notice that we never had to address this approximation in deriving Eq. (2.14), therefore we can expect that an estimator of Eq. (2.17) will differ from $\hat{H}_2(X)$ in Eq. (2.13), i.e. it will have different bias and variance. Finally, substituting the Parzen window estimator of Eq. (2.11) in Eq. (2.17) and rearranging terms, we obtain a nonparametric plug-in estimator for Renyi's α entropy as

$$\hat{H}_\alpha(X) = \frac{1}{1-\alpha}\log\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x_j - x_i)\right)^{\alpha-1} = \frac{1}{1-\alpha}\log(\hat{V}_{\alpha,\sigma}(X)), \qquad (2.18)$$

where the α information potential estimator (the dependence on σ is normally omitted)

$$\hat{V}_{\alpha,\sigma}(X) = \frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(x_j - x_i)\right)^{\alpha-1}.$$

The nonparametric estimator in Eq. (2.18) can still be written as the log of the α-norm of $\hat{V}_\alpha(X) = IP_\alpha(X)$, but again it differs from the IP of Eq. (2.14).. For α = 1, $\hat{V}_1 = 1$ for any PDF. For all $\alpha \geq 0, \alpha \neq 1$ it is a general-purpose estimator and can be used to evaluate α entropy directly from samples or to adapt the weights of a learning system based on an entropic performance index. We now study its properties in detail.

2.4. Properties of Renyi's Nonparametric Entropy Estimators

In the following, all kernel functions and random variable samples are assumed to be single-dimensional unless noted otherwise. The generalization of these results to multidimensional cases is trivial and the proofs follow similar lines. We start by analyzing the accuracy of the approximation of the expected value by the sample average.

Property 2.1: For the special case of Gaussian kernels, the estimator $\hat{V}_\alpha(X)$ of Eq. (2.18) only differs from the $\hat{V}_2(X)$ of Eq. (2.14) by a factor of $\sqrt{2}$ in the kernel size.
Proof: A direct comparison proves the property. This difference stems from the need to approximate the expected value by the sample mean in Eq. (2.18). In fact, the estimator of Eq. (2.18) requires two approximations, the approximation of the expected value by the sample mean and the kernel approximation of the PDF, whereas Eq. (2.13) only requires the kernel approximation of the PDF. Therefore in general they yield two different estimators of the same statistical quantity. However, what is interesting is that for α = 2 the sample mean approximation for finite *N* and Gaussian kernels can still be exactly compensated by a change of the kernel size from σ to $\sigma\sqrt{2}$ in Eq. (2.18).

Property 2.2: For any kernel function κ(*x*) that obeys the relation

$$\kappa^{new}(x_j - x_i) = \int_{-\infty}^{\infty}\kappa^{old}(x - x_i)\cdot\kappa^{old}(x - x_j)dx, \qquad (2.19)$$

where κ^{new}(.) denotes the kernel function used in Eq. (2.18) and κ^{old}(.) denotes the kernel function used in Eq. (2.14), the estimator of Renyi's quadratic entropy of Eq. (2.18) matches the estimator of Eq. (2.14) using the IP.

Proof: Direct substitution proves the property. These properties reiterate the privileged place of the Gaussian kernel and quadratic Renyi's entropy in ITL. The case α = 2 also allows a very interesting link between ITL and kernel learning, and explains the reason why the sample mean approximation is not necessary in IP computation.

It is possible to show using the properties of the Gaussian kernel that Renyi's α entropy estimator can be written exactly as

$$\hat{H}_\alpha(X)=\frac{1}{1-\alpha}\log\left\{\frac{1}{N^\alpha}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{\alpha-1}\left(\frac{1}{\sqrt{\alpha}}\right)\left(\sqrt{2\pi}\sqrt{\alpha\sigma}\right)^{\binom{\alpha}{2}}\sum_{i_1=1}^{N}\cdots\sum_{i_\alpha=1}^{N}\left[\prod_{p=1}^{\alpha}\prod_{q=1,q>p}^{\alpha}G_{\sigma\sqrt{\alpha}}(x_{i_p}-x_{i_q})\right]\right\} \quad (2.20)$$

or

$$\hat{H}_\alpha(X)=\frac{1}{1-\alpha}\log\left\{\frac{1}{N^\alpha}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{\alpha-2}\sum_{i_1=1}^{N}\cdots\sum_{i_\alpha=1}^{N}G_{\sqrt{\alpha}\sigma}\left(\sum_{p=1}^{\alpha}\sum_{q=1,q>p}^{\alpha}\left(x_{i_p}-x_{i_q}\right)^2\right)\right\}. \quad (2.21)$$

In either form one still sees the kernel size for the Gaussian being multiplied by $\sigma\sqrt{\alpha}$ as could be expected from Property 2.1, however, these expressions are not easily compared with Eq. (2.18) even when the kernel is a Gaussian. Therefore, the practical estimation of α Renyi's entropy with kernels will follow the approximation of the expected value by the sample mean as indicated in Eq. (2.18).

Property 2.3. The kernel size must be a parameter that satisfies the scaling property $\kappa_{c\sigma}(x)=\kappa_\sigma(x/c)/c$ for any positive factor *c* [241].

This regulatory condition guarantees that changes in the kernel size can be compensated by linear scaling in the domain and range of the estimated quantities. In the analysis of the eigenstructure of the entropy cost function near the global optimum and in obtaining scale-invariant entropy-based cost functions, this property becomes useful.

Property 2.4. The entropy estimator in Eq. (2.18) is invariant to the mean of the underlying density of the samples as is the actual entropy [86].

Proof. Consider two random variables X and $\bar{X}$ where $\bar{X}=X+m$ with m being a real number. The entropy of $\bar{X}$ becomes

$$\begin{aligned}H_\alpha(\bar{X})&=\frac{1}{1-\alpha}\log\int p_{\bar{X}}^{\alpha}(\bar{x})d\bar{x}=\frac{1}{1-\alpha}\log\int p_{X}^{\alpha}(\bar{x}-m)d\bar{x}\\&=\frac{1}{1-\alpha}\log\int p_{X}^{\alpha}(x)dx=H_\alpha(X).\end{aligned} \quad (2.22)$$

Let $\{x_1, \ldots, x_N\}$ be samples of X, then samples of $\bar{X}$ are $\{x_1+m, \ldots, x_N+m\}$. Therefore, the estimated entropy of $\bar{X}$ is

$$\begin{aligned}\hat{H}_\alpha(\bar{X})&=\frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(\bar{x}_j-\bar{x}_i)\right)^{\alpha-1}\\&=\frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(x_j+m-x_i-m)\right)^{\alpha-1}=\hat{H}_\alpha(X).\end{aligned} \quad (2.23)$$

Due to this property of the entropy and its estimator, when the entropy cost is utilized in supervised learning the mean of the error signal is not necessarily zero, which is a requirement for most applications. This requirement has to be implemented by adding a bias term to the system output that makes the mean of the error equal to zero. Because of this feature, entropy does not define a metric in the space of the samples. We address this point in more detail in Chapter 3. However, when we are interested in the statistical properties of the signals other than their means, this is not a problem.

Property 2.5. The limit of Renyi's entropy as $\alpha \to 1$ is Shannon's entropy. The limit of the entropy estimator in Eq. (2.18) as $\alpha \to 1$ is Shannon's entropy estimated using Parzen windowing with the expectation approximated by the sample mean.

Proof. Notice that Renyi's entropy in Eq. (2.16) is discontinuous at $\alpha = 1$. However, when we take its limit as this parameter approaches one, we get Shannon's entropy as shown in Eq. (2.24) for continuous variables (similarly to Eq. (2.8)),

$$\begin{aligned}\lim_{\alpha\to1} H_\alpha(X) &= \lim_{\alpha\to1}\frac{1}{1-\alpha}\log\int p_X^\alpha(x)dx = \frac{\lim\limits_{\alpha\to1}\int \log p_X(x)\cdot p_X^\alpha(x)dx \Big/ \int p_X^\alpha(x)dx}{\lim\limits_{\alpha\to1} -1} \\ &= -\int p_X(x)\cdot\log p_X(x)dx = H_S(X)\end{aligned} \tag{2.24}$$

The derivation of this result for the estimator in Eq. (2.18) is shown in Eq. (2.25).

$$\begin{aligned}\lim_{\alpha\to1}\hat{H}_\alpha(X) &= \lim_{\alpha\to1}\frac{1}{1-\alpha}\log\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x_j-x_i)\right)^{\alpha-1} \\ &= -\lim_{\alpha\to1}\frac{\left(\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x_j-x_i)\right)^{\alpha-1}\log\left(\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x_j-x_i)\right)\right)}{\left(\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x_j-x_i)\right)^{\alpha-1}\right)} \\ &= \frac{-1}{N}\sum_{j=1}^{N}\log\left(\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x_j-x_i)\right) = \hat{H}_S(X).\end{aligned} \tag{2.25}$$

In terms of adaptation, this means that all the conclusions drawn in this research about Renyi's entropy, its estimator, and training algorithms based on Renyi's entropy apply in the limit of $\alpha \to 1$, to Shannon's definition as well.

Property 2.6. In order to maintain consistency with the scaling property of the actual entropy, if the entropy estimate of samples $\{x_1, \dots, x_N\}$ of a random variable X is estimated using a kernel size of σ, the entropy estimate of the samples $\{cx_1, \dots, cx_N\}$ of a random variable cX must be estimated using a kernel size of $|c|\sigma$.

Proof. Consider the Renyi's entropy of the random variable cX, whose PDF is $p_X(x/c)/|c|$ in terms of the PDF of the random variable X and the scaling coefficient c.

$$H_\alpha(cX) = \frac{1}{1-\alpha}\log\int_{-\infty}^{\infty}\frac{1}{|c|^\alpha}\,p_X^\alpha(\frac{x}{c})dx = H_\alpha(X)+\log|c|. \tag{2.26}$$

Now consider the entropy estimate of the samples $\{cx_1, \ldots, cx_N\}$ using the kernel size $|c|\sigma$.

$$\begin{aligned}
\hat{H}_\alpha(cX) &= \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_{|c|\sigma}(\frac{x_j - x_i}{c})\right)^{\alpha-1} \\
&= \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\frac{1}{|c|}\kappa_{\sigma}(\frac{cx_j - cx_i}{c})\right)^{\alpha-1} \\
&= \frac{1}{1-\alpha}\log\frac{1}{|c|^{\alpha-1}N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_{\sigma}(x_j - x_i)\right)^{\alpha-1} \\
&= \hat{H}_\alpha(X)+\log|c|.
\end{aligned} \tag{2.27}$$

This property is crucial when the problem requires a scale-invariant cost function as in blind deconvolution illustrated in Chapter 8. The scaling of the kernel size as described above according to the norm of the weight vector guarantees that the nonparametric estimation of the scale-invariant cost function possesses this property as well.

Property 2.7. When estimating the joint entropy of an n-dimensional random vector X from its samples $\{x_1, \ldots, x_N\}$, use a multidimensional kernel that is the product of single-dimensional kernels. In this way, the estimate of the joint entropy and estimate of the marginal entropies are consistent.

Proof. Let the random variable X_o be the oth component of X. Consider the use of single-dimensional kernels $\kappa_{\sigma_o}(.)$ for each of these components. Also assume that the multidimensional kernel used to estimate the joint PDF of X is $\kappa_\Sigma(.)$. The Parzen estimate of the joint PDF is then given by

$$\hat{p}_X(x) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\Sigma(x-x_i). \tag{2.28}$$

Similarly, the Parzen estimate of the marginal density of X_o is

$$\hat{p}_{X,o}(x) = \frac{1}{N}\sum_{i=1}^{N}\kappa_{\sigma_o}(x_o - x_o(i)). \tag{2.29}$$

Without loss of generality, consider the marginal PDF of X_1 derived from the estimate of the joint PDF in Eq. (2.28).

$$\hat{p}_{X,1}(x_1) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\bar{p}_X(x_1,\ldots,x_n)dx_2,\ldots,dx_n = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{1}{N}\sum_{i=1}^{N}\kappa_{\Sigma}(x_1-x_1(i),\ldots,x_n-x_n(i))dx_2,\ldots,dx_n$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\kappa_{\Sigma}(x_1-x_1(i),\ldots,x_n-x_n(i))dx_2,\ldots,dx_n. \qquad (2.30)$$

Now, assuming that the joint kernel is the product of the marginal kernels evaluated at the appropriate values (i.e., $\kappa_{\Sigma}(x) = \prod_{o=1}^{N}\kappa_{\sigma_o}(x_o)$), we get Eq. (2.31). Thus, this choice of the multidimensional kernel for joint entropy estimation guarantees consistency between the joint and marginal PDF and entropy estimates. This property is, in fact, critical for the general PDF estimation problem besides being important in entropy estimation.

$$p_{X,1}(x_1) = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{o=1}^{n}\kappa_{\sigma_o}(x_o-x_o(i))dx_2,\ldots,dx_n$$

$$= \frac{1}{N}\sum_{i=1}^{N}\kappa_{\sigma_1}(x_1-x_1(i))\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{o=2}^{n}\kappa_{\sigma_o}(x_o-x_o(i))dx_2,\ldots,dx_n \qquad (2.31)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\kappa_{\sigma_1}(x_1-x_1(i))\left(\prod_{o=2}^{n}\int_{-\infty}^{\infty}\kappa_{\sigma_o}(x_o-x_o(i))dx^o\right) = \hat{p}_{X,1}(x_1).$$

This important issue should be considered in adaptation scenarios where the marginal entropies of multiple signals and their joint entropy are used in the cost function simultaneously. It is desirable to have consistency between the marginal and joint entropy estimates.

Theorem 2.1. *The entropy estimator in Eq.* (2.18) *is consistent if the Parzen windowing and the sample mean are consistent for the actual PDF of the i.i.d. samples.*

Proof. The proof follows immediately from the consistency of the Parzen window estimate for the PDF and the fact that as N goes to infinity the sample mean converges to the expected value which makes Eq. (2.18) approach Eq. (2.16) (e.g., the sample mean estimate is not consistent for infinite-variance PDFs).

This theorem is important because it points out the asymptotic limitations of the estimator. In adaptation and learning from finite samples, because we rarely have huge datasets, consistency is not the primary issue, but the bias and variance of the estimator must still be known. Their effect in the location of the extremum of the function in the space of the parameters is the real issue.

Theorem 2.2. *If the maximum value of the kernel* $\kappa_\sigma(\xi)$ *is achieved when* $\xi = 0$, *then the minimum value of the entropy estimator in Eq.* (2.18) *is achieved when all samples are equal to each other, that is,* $x_1 = \ldots = x_N = c$ [86].

Proof. By substitution, we find that the entropy estimator takes the value $-\log\kappa_\sigma(0)$ when all samples are equal to each other. We need to show that

$$\frac{1}{1-\alpha}\log\frac{1}{N^{\alpha}}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_{\sigma}(x_j-x_i)\right)^{\alpha-1}\geq-\log\kappa_{\sigma}(0)\,. \tag{2.32}$$

For $\alpha>1$, this is equivalent to showing that

$$\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_{\sigma}(x_j-x_i)\right)^{\alpha-1}\leq N^{\alpha}\kappa_{\sigma}^{\alpha-1}(0)\,. \tag{2.33}$$

Replacing the left-hand side of Eq. (2.33) with its upper bound we get Eq. (2.34). Because the kernel function is chosen such that its maximum occurs when its argument is zero, we obtain the desired result given in Eq. (2.33).

$$\begin{aligned}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_{\sigma}(x_j-x_i)\right)^{\alpha-1}&\leq N\max_{j}\left[\left(\sum_{i=1}^{N}\kappa_{\sigma}(x_j-x_i)\right)^{\alpha-1}\right]\\ &\leq N\max_{j}\left[N^{\alpha-1}\max_{i}\underset{i}{\kappa_{\sigma}^{\alpha-1}(x_j-x_i)}\right]=N^{\alpha}\max_{i,j}\kappa_{\sigma}^{\alpha-1}(x_j-x_i).\end{aligned} \tag{2.34}$$

The proof for the case $\alpha<1$ is similar. It uses the min operator instead of max due to the direction of the inequality.

In supervised training, it is imperative that the cost function achieve its global minimum when all the error samples are zero. Minimum error entropy learning using this entropy estimator, which is introduced in Chapter 3, becomes a valid supervised training approach with this property of the entropy estimator. In addition, the unsupervised training scenarios such as minimum entropy blind deconvolution, which are discussed in Chapter 8, benefit from this property of the estimator as well.

Theorem 2.3. *If the kernel function $\kappa_\sigma(.)$ is continuous, differentiable, symmetric, and unimodal, then the global minimum described in Theorem* 2.2 *of the entropy estimator in Eq.* (2.18) *is smooth, that is, it has a zero gradient and a positive semidefinite Hessian matrix.*

Proof. Let $\bar{\mathbf{x}}=[x_1\ \ ,\ldots,\ \ x_N]^T$ be the data samples collected in a vector for notational simplicity. Without loss of generality, consider the dataset given by $\bar{\mathbf{x}}=0$, meaning all samples are zero. With some algebra, the gradient and the Hessian matrix of the expression in Eq. (2.18) with respect to $\bar{\mathbf{x}}$ are found as

$$\begin{aligned}\frac{\partial\hat{H}_{\alpha}}{\partial x_k}&=\frac{1}{1-\alpha}\frac{\partial\hat{V}_{\alpha}/\partial x_k}{\hat{V}_{\alpha}}\\ \frac{\partial^2\hat{H}_{\alpha}}{\partial x_l\partial x_k}&=\frac{1}{1-\alpha}\frac{(\partial^2\hat{V}_{\alpha}/\partial x_l\partial x_k)\hat{V}_{\alpha}-(\partial\hat{V}_{\alpha}/\partial x_k)(\partial\hat{V}_{\alpha}/\partial x_l)}{\hat{V}_{\alpha}^2}.\end{aligned} \tag{2.35}$$

where the variable $\hat{V}_{\alpha}$ is the argument of the logarithm in the final expression in Eq. (2.18). Evaluating these expressions at $\bar{\mathbf{x}}=0$, which corresponds to the maximum value of the kernel we get

$$
\begin{aligned}
\hat{V}_\alpha\Big|_{\bar{x}=0} &= \kappa_\sigma^{\alpha-1}(0) \\
\frac{\partial \hat{V}_\alpha}{\partial x_k}\Big|_{\bar{x}=0} &= \frac{(\alpha-1)}{N^\alpha}\left[N^{\alpha-1}\kappa_\sigma^{\alpha-2}(0)\kappa'(0) - N^{\alpha-1}\kappa_\sigma^{\alpha-2}(0)\kappa'(0)\right] = 0 \\
\frac{\partial^2 \hat{V}_\alpha}{\partial x_k^2}\Big|_{\bar{x}=0} &= \frac{(\alpha-1)(N-1)\kappa_\sigma^{\alpha-3}(0)}{N^2}\left[(\alpha-2)\kappa'^2(0) + 2\kappa(0)\kappa''(0)\right] \\
\frac{\partial^2 \hat{V}_\alpha}{\partial x_l\,\partial x_k}\Big|_{\bar{x}=0} &= -\frac{(\alpha-1)\kappa_\sigma^{\alpha-3}(0)}{N^2}\left[(\alpha-2)\kappa'^2(0) + 2\kappa(0)\kappa''(0)\right],
\end{aligned}
\tag{2.36}
$$

which shows that the gradient vector is zero and that the Hessian matrix is composed of

$$
\frac{\partial^2 \hat{H}_\alpha}{\partial x_l \partial x_k}\Big|_{\bar{x}=0} = \begin{cases} -(N-1)\kappa_\sigma^{-\alpha-1}(0)\left[(\alpha-2)\kappa'^2(0) + 2\kappa(0)\kappa''(0)\right]/N^2, & l = k \\ \kappa_\sigma^{-\alpha-1}(0)\left[(\alpha-2)\kappa'^2(0) + 2\kappa(0)\kappa''(0)\right]/N^2, & l \neq k \end{cases}
\tag{2.37}
$$

Denoting the diagonal terms by *a* and the off-diagonal terms by *b*, we can determine all the eigenvalue-eigenvector pairs of this matrix to be

$$\{0,[1,...,1]^T\},\{aN/(N-1),[1,-1,0,...,0]^T\},\{aN/(N-1),[1,0,-1,0,...,0]^T\},...$$

Notice that the nonzero eigenvalue has a multiplicity of $N-1$ and for a kernel function as described in the theorem and for $N > 1$ this eigenvalue is positive, because the kernel evaluated at zero is positive, the first derivative of the kernel evaluated at zero is zero, and the second derivative is negative. Thus the Hessian matrix at the global minimum of the entropy estimator is negative semidefinite. This is to be expected because there is one eigenvector corresponding to the direction that only changes the mean of data, along which the entropy estimator is constant due to Property 2.4.

In adaptation using numerical optimization techniques, it is crucial that the global optimum be a smooth point in the weight space with zero gradient and finite-eigenvalue Hessian. This last theorem shows that the nonparametric estimator is suitable for entropy minimization adaptation scenarios.

Property 2.8. If the kernel function satisfies the conditions in Theorem 2.3, then in the limit, as the kernel size tends to infinity, the quadratic entropy estimator approaches the logarithm of a scaled and biased version of the sample variance.

Proof. Let $\{x_1,\ldots,x_N\}$ be the samples of X. We denote the second-order sample moment and the sample mean with the following.

$$\overline{x^2} = \frac{1}{N}\sum_{j=1}^{N} x_j^2 \qquad \bar{x}^2 = \left(\frac{1}{N}\sum_{j=1}^{N} x_j\right)^2 .$$

By assumption the kernel size is very large, therefore the pairwise differences of samples will be very small compared to the kernel size, thus allowing the second-order Taylor series expansion of the kernel function around zero to be a valid approximation.

Also, due to the kernel function being symmetric and differentiable, its first-order derivative at zero will be zero yielding

$$\kappa_\sigma(\xi) \approx \kappa_\sigma(0) + \kappa'_\sigma(0)\xi + \kappa''_\sigma(0)\xi^2/2 = \kappa_\sigma(0) + \kappa''_\sigma(0)\xi^2/2\,. \tag{2.38}$$

Substituting this in the quadratic entropy estimator obtained from Eq. (2.18) by substituting $\alpha = 2$, we get Eq. (2.39), where $\overline{x^2} - \bar{x}^2$ is the sample variance. Notice that the kernel size affects the scale factor multiplying the sample variance in Eq. (2.39). In addition to this, there is a bias depending on the kernel's center value.

$$\begin{aligned}\hat{H}_2(X) &\approx -\log\left[\frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\left(\kappa_\sigma(0) + \kappa''_\sigma(0)(x_j - x_i)^2/2\right)\right] \\ &= -\log\left[\kappa_\sigma(0) + \frac{1}{2}\kappa''_\sigma(0)\left(\frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\left(x_j^2 - 2x_j x_i + x_i^2\right)\right)\right] \\ &= -\log\left[\kappa_\sigma(0) + \frac{1}{2}\kappa''_\sigma(0)\left(\overline{x^2} - \bar{x}^2\right)\right].\end{aligned} \tag{2.39}$$

<u>Property 2.9</u>. In the case of joint entropy estimation, if the multidimensional kernel function satisfies $\kappa_\Sigma(\xi) = \kappa_\Sigma(R^{-1}\xi)$ for all orthonormal matrices *R*, then the entropy estimator in Eq. (2.18) is invariant under rotations as is the actual entropy of a random vector *X*. Notice that the condition on the joint kernel function requires hyperspherical symmetry.

<u>Proof.</u> Consider two *n*-dimensional random vectors *X* and $\bar{X}$ related to each other with $\bar{X} = RX$ where *R* is an *n* × *n* real orthonormal matrix. Then the entropy of $\bar{X}$ is

$$\begin{aligned}H_\alpha(\bar{X}) &= \frac{1}{1-\alpha}\log\int_{-\infty}^{\infty} p_{\bar{X}}^{\alpha}(\bar{x})d\bar{x} = \frac{1}{1-\alpha}\log\int_{-\infty}^{\infty}\frac{1}{|R|^{\alpha}}\,p_X^{\alpha}(R^{-1}\bar{x})d\bar{x} \\ &= \frac{1}{1-\alpha}\log\int_{-\infty}^{\infty}\frac{1}{|R|^{\alpha}}\,p_X^{\alpha}(x)|R|dx = \frac{1}{1-\alpha}\log|R|^{1-\alpha}\int_{-\infty}^{\infty}p_X^{\alpha}(x)dx \\ &= H_\alpha(X) + \log|R| = H_\alpha(X).\end{aligned} \tag{2.40}$$

Now consider the estimation of the joint entropy of $\bar{X}$ from its samples, which are given by $\{Rx_1,\ldots,Rx_N\}$, where $\{x_1,\ldots,x_N\}$ are samples of *X*. Suppose we use a multidimensional kernel $\kappa_\Sigma(.)$ that satisfies the required condition. This results in Eq. (2.41). In adaptation scenarios where the invariance-under-rotations property of entropy needs to be exploited, the careful choice of the joint kernel becomes important. Property 2.5 describes how to select kernel functions in such situations.

$$\begin{aligned}\hat{H}_\alpha(\bar{X}) &= \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\Sigma(Rx_j - Rx_i)\right)^{\alpha-1}\\ &= \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\frac{1}{|R|}\kappa_\Sigma(R^{-1}(Rx_j - Rx_i))\right)^{\alpha-1}\\ &= \frac{1}{1-\alpha}\log|R|^{\alpha-1}\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\Sigma(x_j - x_i)\right)^{\alpha-1}\\ &= \hat{H}_\alpha(X)\end{aligned} \tag{2.41}$$

<u>Theorem 2.4.</u> $\lim_{N\to\infty}\hat{H}_\alpha(X) = H_\alpha(\hat{X}) \ge H_\alpha(X)$, *where* $\hat{X}$ *is a random variable with the PDF* $f_X(.) * \kappa_\sigma(.)$. *The equality occurs if and only if the kernel size is zero. This result is also valid on the average for the finite-sample case.*

<u>Proof.</u> It is well known that the Parzen window estimate of the PDF of X converges consistently to $f_X(.) * \kappa_\sigma(.)$. Therefore, the entropy estimator in Eq. (2.18) converges to the actual entropy of this PDF. To prove the inequality consider

$$e^{(1-\alpha)H_\alpha(\hat{X})} = \int_{-\infty}^{\infty} p_{\hat{X}}^{\alpha}(y)dy = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\kappa_\sigma(\tau)p_X^\alpha(y-\tau)d\tau\right]dy. \tag{2.42}$$

Using Jensen's inequality for convex and concave cases, we get Eq. (2.43), where we defined the mean-invariant quantity $V_\alpha(X)$ as the integral of the αth power of the PDF of X, which is the argument of the log in the definition of Renyi's entropy given in Eq. (2.16). Reorganizing the terms in Eq. (2.43) and using the relationship between entropy and information potential, regardless of the value of α and the direction of the inequality, we arrive at the conclusion $H_\alpha(\hat{X}) \ge H_\alpha(X)$. The fact that these results are also valid on the average for the finite-sample case is due to the property $E[\hat{p}_X(.)] = p_X(.) * \kappa_\sigma(.)$ of Parzen windowing, which relates the average PDF estimate to the actual value and the kernel function.

$$\begin{aligned}\exp((1-\alpha)H_\alpha(\hat{X})) &\overset{\alpha>1}{\le}\left(\overset{\alpha<1}{\ge}\right)\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\kappa_\sigma(\tau)[p_X(y-\tau)]^\alpha d\tau\right]dy\\ &= \int_{-\infty}^{\infty}\kappa_\sigma(\tau)\left[\int_{-\infty}^{\infty}[p_X(y-\tau)]^\alpha dy\right]d\tau = \int_{-\infty}^{\infty}\kappa_\sigma(\tau)V_\alpha(X)d\tau\\ &= V_\alpha(X)\cdot\int_{-\infty}^{\infty}\kappa_\sigma(\tau)d\tau = V_\alpha(X).\end{aligned} \tag{2.43}$$

This theorem is useful in proving asymptotic noise rejection properties of the entropy-based adaptation criteria, and shows that for entropy minimization, the proposed estimator provides a useful approximation in the form of an upper bound to the true entropy of the signal under consideration.

2.5. Bias and Variance of the Information Potential Estimator

IP Estimator Bias

In this section, the bias of the information potential is analyzed for finite samples using the shape of the data probability density function (which is unknown for most cases, but provides understanding of the factors involved). We call the attention of the readers to the analysis of the density estimation in Appendix A, which should be used to contrast the results obtained in this section. The same basic approach is taken here, but a simplified notation will be used and some steps are omitted. We choose the Gaussian kernel for density estimation. The information potential estimator is

$$\hat{V}_{2,\sigma}(X) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_\sigma(x_j - x_i)\ . \tag{2.44}$$

The IP bias is obtained by taking the expectation

$$E[\hat{V}_2(X)] = E[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_\sigma(x_j - x_i)] = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} E[G_\sigma(x_j - x_i)] = E[G_\sigma(x_j - x_i)]. \tag{2.45}$$

Now expanding the PDF in Taylor series and using the i.i.d. assumption on the data and the definition of expected value

$$\begin{aligned}
E[G_\sigma(x_j - x_i)] &= \int\int G_\sigma(x_i - x_j)p(x_i)p(x_j)dx_i dx_j \\
&= \int p(x_i)[\int G_\sigma(s)p(x_i + \sigma\ s)ds]dx_i \qquad s = (x_i - x_j)/\sigma \\
&= \int p(x)\{\int G_\sigma(s)[p(x) + \sigma\ sp(x) + 1/2(\sigma^2 s^2 p''(x) + o(\sigma^4))]ds\}dx \\
&\approx \int p(x)[p(x) + 1/2(\sigma^2 p''(x))\mu_2(G)]dx \qquad as \qquad \sigma \to 0 \\
&= \int p^2(x)dx + (\sigma^2/2)\mu_2(G)\int p(x)p''(x)dx.
\end{aligned} \tag{2.46}$$

where for the Gaussian kernel $\mu_2(G) = \int \sigma^2 G_\sigma(s)ds = 1$. Now

$$\int p(x)p''(x)dx = E[p''(X)] \tag{2.47}$$

and combining the above equations we obtain

$$Bias[\hat{V}_2(X)] = E[\hat{V}_2(X)] - \int p^2(x)dx = (\sigma^2/2)E[p''(X)] \tag{2.48}$$

We see that the bias of IP for Gaussian kernels increases proportionally to the square of the kernel size multiplied by the expected value of the PDF second derivative. This result has the same basic form of Eq. (2.12) for kernel density estimation (see Appendix A), except that the double derivative of the PDF is substituted by its expected value, which is better behaved.

IP Estimator Variance

To analyze the variance of the information potential, we rewrite Eq. (2.44) as

$$N^2\hat{V}(X) = \sum_i \sum_j G_\sigma(x_i - x_j)$$

and take the variance of both sides to obtain:

$$\begin{aligned} N^4 Var(\hat{V}(X)) &= N^4 \left\lfloor E(\hat{V}^2(X)) - (E(\hat{V}(X)))^2 \right\rfloor \\ &= \sum_i \sum_j \sum_k \sum_l \left\{ E\left[G_\sigma(x_i - x_j) G_\sigma(x_k - x_l)\right] - E\left[G_\sigma(x_i - x_j)\right] E\left[G_\sigma(x_k - x_l)\right] \right\} \end{aligned} \tag{2.49}$$

The right-hand side of Eq. (2.49) consists of N^4 terms. These terms can be classified into four categories according to the possible values of *i, j, k,* and *l*:

1. If *i, j, k, l* are all different among each other, then, according to the independence assumption, their joint distribution can be factorized as

$$E\left\lfloor G_\sigma(x_i - x_j) G_\sigma(x_k - x_l)\right\rfloor = E\left\lfloor G_\sigma(x_i - x_j)\right\rfloor E\left[G_\sigma(x_k - x_l)\right] \tag{2.50}$$

therefore, all the positive terms in the summation cancel out the corresponding negative terms.

2. If $j \neq i = k \neq l$ and $j \neq l$, then, x_j, x_l would be independent when $x_i = x_k$ and those terms can be calculated as

$$E\left\lfloor G_\sigma(x_i - x_j) G_\sigma(x_l - x_i)\right\rfloor - E\left\lfloor G_\sigma(x_i - x_j)\right\rfloor E\left[G_\sigma(x_k - x_l)\right] , \tag{2.51}$$

Choosing different values of *i, j,* and *k* there are totally $N(N-1)(N-2)$ terms like Eq. (2.51).

3. If $i = k \neq j = l$, by the same argument, there are $N(N-1)$ terms and these terms are all equal:

$$\begin{aligned} &E\left\lfloor G_\sigma(x_i - x_j) G_\sigma(x_j - x_i)\right\rfloor - E\left\lfloor G_\sigma(x_i - x_j)\right\rfloor E\left\lfloor G_\sigma(x_j - x_i)\right\rfloor \\ &= E\left[G_\sigma(x_i - x_j)^2\right] - E\left[G_\sigma(x_i - x_j)\right]^2 = Var[G_\sigma(x_i - x_j)]. \end{aligned} \tag{2.52}$$

4. If $i = k = j \neq l$, then, similarly, the independent terms can be written as:

$$\begin{aligned} &E\left[G_\sigma(x_i - x_i) G_\sigma(x_i - x_l)\right] - E\left[G_\sigma(x_i - x_i)\right] E\left[G_\sigma(x_i - x_l)\right] \\ &= E\left[G_\sigma(0) G_\sigma(x_i - x_l)\right] - E\left[G_\sigma(0) G_\sigma(x_i - x_l)\right] = 0. \end{aligned} \tag{2.53}$$

From this discussion, we see that only Cases 2 and 3 will yield a nonzero value, and they will affect the variance of the information potential with different weights; that is, since the number of terms in Case 2 is $N(N-1)(N-2)$ which is proportional to N^3 while the number of terms in Case 3 is $N(N-1)$ which is proportional to N^2. Thus, as the number of sample *N* increases, Eq. (2.51) becomes dominant. We denote

$$\begin{aligned} a &= E\left\lfloor K_\sigma(x_i - x_j) K_\sigma(x_j - x_l)\right\rfloor - E\left\lfloor K_\sigma(x_i - x_j)\right\rfloor E\left\lfloor K_\sigma(x_j - x_l)\right\rfloor \\ b &= Var[K_\sigma(x_i - x_j)], \end{aligned} \tag{2.54}$$

where *K* is the Gaussian kernel of Eq. (1.50) without the division by σ. If $a \neq 0$ which is generally true for most probability density functions we can write:

$$Var(\hat{V}(X)) = E(\hat{V}^2(X)) - (E(\hat{V}(X)))^2 = \frac{aN(N-1)(N-2)+bN(N-1)}{\sigma N^4} \approx \frac{a}{N\sigma}, \qquad N \to \infty. \tag{2.55}$$

So, from this analysis, we conclude that the variance of the information potential will decrease inversely proportional to *N*, which is a comfortable result for estimation. The asymptotic mean integrated square error (AMISE) of the IP is therefore

$$AMISE(\hat{V}(X)) = E[\int (\hat{V}(X) - V(X))^2 dx] = \frac{\sigma^4}{2}\int (p''(x))^2 dx + \frac{aN(N-1)(N-2)+bN(N-1)}{\sigma N^4} \tag{2.56}$$

Notice that the AMISE will tend to zero when the kernel size goes to zero and the number of samples goes to infinity with $N\sigma \to \infty$, that is, the IP is a consistent estimator of the 2-norm of the PDF. Unlike the estimators for the mean and variance, the IP is a biased estimator of the 2-norm of the PDF for finite bandwidth kernels. If we compare closely Eqs. (2.49), (2.55), and (2.56) with the well-known Parzen estimation (Eq. (2.12) and [300]) we see that they are assymptotically the same; the bias is proportional to σ^2 and the variance decreases proportionally to $N\sigma$. The similarity of Eq. (2.56) with kernel density estimation shows that the body of knowledge in density estimation is directly applicable to the estimation of IP, or that the IP is essentially a kernel estimator of the 2-norm of the PDF. We also can conclude that the estimators of Eqs. (2.14) and (2.18) for the quadratic Renyi's entropy trade bias with variance; that is Eq. (2.14) has larger bias but smaller variance.

2.6. Physical Interpretation of Renyi's Entropy Kernel Estimators

There is a useful physical analogy for the kernel estimator in Renyi's entropy as defining an *information potential field* [86]. This analogy has its roots in the link between Renyi's entropy and the norms of the PDF of the data. Indeed, because the kernels in PDF estimation are positive functions that decay with the distance between samples, one can think that one kernel placed on a sample creates a potential field in the sample space, just as physical particles create a gravity field in space. However, in our case the law of interaction is dictated by the kernel shape. The density of samples is measured by the PDF, therefore the potential field in the space of the samples is an approximation of the PDF shape. In this context, samples can be named *information particles* and they interact in the information potential field of the PDF creating *information forces* [86]. The only difference is that this framework must obey the sum constraint of PDFs (so sums are replaced by averages). We explain these concepts next.

Consider $\hat{V}_2(X)$ in Eq. (2.14) as the average sum of interactions from each sample x_j in the sample set; that is, $\hat{V}_2(X) = 1/N\sum_{j=1}^{N}\hat{V}_2(x_j)$ where

$$\hat{V}_2(x_j) \overset{\Delta}{=} 1/N\sum_{i=1}^{N}\hat{V}_2(x_j;x_i) \qquad and \qquad \hat{V}_2(x_j;x_i) = G_{\sigma\sqrt{2}}(x_j - x_i) \tag{2.57}$$

which basically measures the effect of the potential field in the space location occupied by the sample x_j due to all the other samples x_i. The sample-by-sample interaction Eq. (2.57) is controlled as we can see by the kernel used in the analysis. The analogy with fields is accurate if we think of the "average" field produced by the samples and this is required to establish the link to PDFs which must add to one. $\hat{V}_2(x_j)$ can be recognized as the value of the PDF estimated at x_j so $\hat{V}_2(x)$, the estimated PDF with kernels for an arbitrary point x in the space, can be properly called the information potential field. The IP is just the average value of the information potential field of the sample set (hence the name).

For Gaussian kernels, the derivative of the information potential with respect to the position of sample x_j is easily evaluated as

$$\frac{\partial}{\partial x_j}\hat{V}_2(x_j)=\frac{1}{N}\sum_{i=1}^{N}G'_{\sigma\sqrt{2}}(x_j-x_i)=\frac{1}{2N\sigma^2}\sum_{i=1}^{N}G_{\sigma\sqrt{2}}(x_j-x_i)(x_i-x_j). \qquad (2.58)$$

This expression estimates the information force exerted on sample x_j due to all the other samples. Note that the derivative of the Gaussian evaluated at zero is zero (any kernel that is symmetric, continuous, and maximum at the origin has a similar property). We can also regard Eq. (2.58) as the average contribution of derivatives due to all other samples. Denoting the contribution of a single sample x_i as $F_2(x_j;x_i)$, and the overall derivative with respect to x_j as $F_2(x_j)$, we get

$$\begin{aligned} F_2(x_j;x_i) &\overset{\Delta}{=} G'_{\sigma\sqrt{2}}(x_j-x_i) \\ F_2(x_j) &\overset{\Delta}{=} \frac{\partial}{\partial x_j}\hat{V}_2(x_j)=\frac{1}{N}\sum_{i=1}^{N}F_2(x_j;x_i). \end{aligned} \qquad (2.59)$$

We name these two quantities the *information force on sample x_j due to sample x_i* and the (*total*) *information force acting on sample x_j*, respectively. Figure 2.4 shows one projection of the information force created by one sample at the origin (Gaussian kernel) in 2D space.

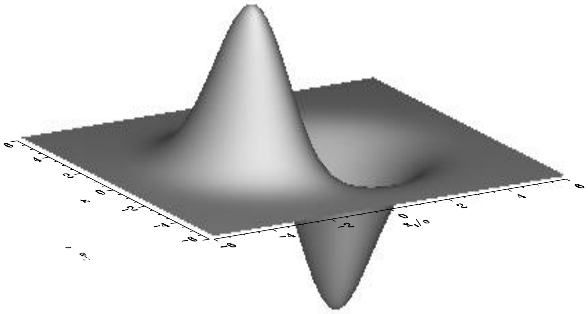


Figure 2.4. The information force created by one information particle placed at the origin in 2D space (Gaussian kernel) in normalized coordinates (x/σ, y/σ) (from [252]).

It is instructive to visualize the procedure for the calculation of the information potential and the information force with a Gaussian kernel. For a dataset $\{x_i\}$ in R^n two matrices can be defined as

$$\begin{cases}\mathbf{D}=\{\mathbf{d}_{ij}\}, & \mathbf{d}_{ij}=\mathbf{x}_i-\mathbf{x}_j\\ \varsigma=\{\hat{V}_{ij}\} & \hat{V}_{ij}=G_{\sigma\sqrt{2}}(\mathbf{d}_{ij}),\end{cases}\tag{2.60}$$

where **D** is a matrix of distances, with vector elements in R^n, and ς a matrix of scalar values where each element quantifies the interaction between two points in the lattice by the kernel, which gives rise to a similarity matrix. From these quantities, all the quantities of information potential field $V(i)$ at location x_i, information force field $F(i)$, and the information potential $V(X)$ can be easily computed for any Parzen kernel. In fact, for the specific case of the Gaussian kernel they are:

$$\text{fields}\begin{cases}\hat{V}(i)=\dfrac{1}{N}\sum_{j=1}^{N}\hat{V}_{i,j}\\ \hat{F}(i)=\dfrac{-1}{2N\sigma^2}\sum_{j=1}^{N}\hat{V}_{i,j}d_{i,j}\end{cases}\tag{2.61}$$

$$\hat{V}(X)=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{V}_{i,j}=\frac{1}{N}\sum_{i=1}^{N}\hat{V}(i).$$

For further reference, notice that information fields are computed with a single sum over columns (or rows) whereas the information potential requires double sums. Because the computation is done with pairs of samples, it can be visualized in a grid where the sample is the axes that work as pointers to the pairwise distances $d_{i,j}$, and the Gaussian is computed directly with this information (Figure 2.5). We can also conclude that the computation complexity of this class of algorithms is $O(N^2)$ or $O(N)$ depending on the quantity of interest, where N is the number of data samples.

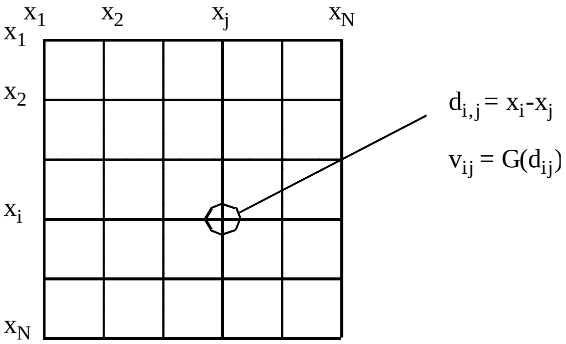
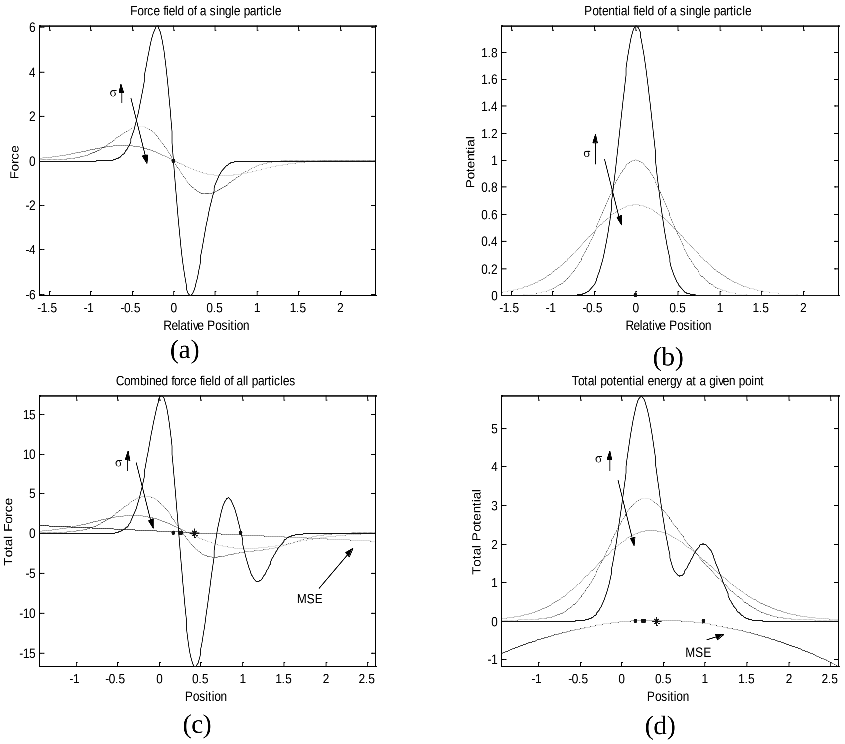


Figure 2.5. The structure of matrix D and ς.

Illustration of Information Forces

Two numerical examples illustrate the information forces and information potentials in single-dimensional and two-dimensional cases. In the first illustration, we consider the single-dimensional case with the kernel function chosen to be a Gaussian. In Figure 2.6, the one-dimensional information forces and information potential fields are shown for various kernel sizes [86]. The attractive force field of an individual particle centered at the origin is plotted in Figure 2.6a. The forces can be made repulsive by introducing a negative sign in the definition. This procedure corresponds to choosing between minimizing or maximizing the sample entropy. Figure 2.6b shows the information potential at any point due to the existence of this particle at the origin as a function of distance to the origin. To further investigate the effect of additional samples on the potential and force fields, we position three additional randomly located samples. The overall quadratic information force field obtained by superposition of the individual forces of these four particles is shown in Figure 2.6c, and the overall quadratic information potential at a given location is presented as a function of position in Figure 2.6d. All plots include illustrations for various values of the selected kernel size. Notice that, as a consequence of the equivalence with sample variance showed in Property 2.4, as the kernel size increases, the effective force becomes a linear function of distance, and is shown with the label *MSE* in Figure 2.6d. For different kernel functions, different force field definitions can be obtained, changing the adaptation dynamics.



(a)

(b)

(c)

(d)

Figure 2.6. Forces and potentials as a function of position for different values of kernel size (a) force due to a single particle; (b) potential due to a single particle; (c) overall quadratic force at a given position due to all particles; (d) total quadratic potential at a given position (from [91]).

As a second illustration, a snapshot of a two-dimensional entropy maximization scenario is depicted in Figure 2.7, where the particles are bounded to within a unit square and interact under the quadratic force definition with a Gaussian kernel. The objective is to maximize the entropy of the sample ensemble, therefore the forces become repulsive and they stabilize in an arrangement that fills the space uniformly with samples. Given a set of randomly spaced samples in the unit square, when the forces acting on each sample are evaluated, it becomes evident that the information particles are pushed by the other particles in order to move along the direction of maximal entropy. Notice also that the forces tend to be larger for samples away from the center of the cluster (the lines attached to each sample are vectors that display intensity and point to the direction of change).

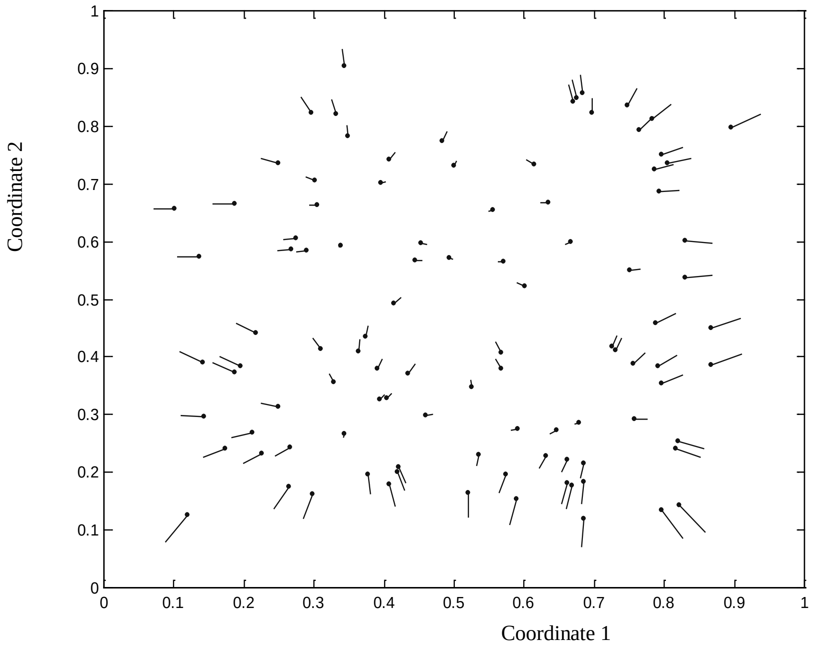


Figure 2.7. A snapshot of the locations of the information particles and the instantaneous quadratic information forces acting on them to maximize the joint entropy in the two-dimensional unit square (from [91]).

Wave Function Interpretation of the Information Potential Estimator

There is another “quantum theory” interpretation of kernel density estimation that is worth presenting [154]. As we have seen, the kernel estimator creates a probability density over the space of the samples. The stationary (time-independent) Schrödinger equation for a particle in the presence of a potential field can be written as

$$\frac{\hbar^2}{2m}\nabla^2\psi(x)+\psi(x)[E-V_Q(x)]=0\,, \tag{2.62}$$

where h is the Plank’s constant, m the mass of the particle, and the wave function ψ determines the spatial probability of the particle with $p(x)=|\psi(x)|^2$. $V_Q(x)$ is the “quantum” potential energy as a function of position, E corresponds to the allowable energy state of the particle, and ψ becomes the corresponding eigenvector. For the set of information particles with the Gaussian kernel, the wavefunction for a set of N, one–dimensional, information particles can be written

$$\psi(x)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}G_\sigma(x-x_i)}.$$

To simplify the derivation and if we are not interested in the physical meaning of the eigenfunctions, we can redefine $\psi(w)$ as

$$\psi(x) = \sum_{i=1}^{N} G_\sigma(x - x_i). \qquad (2.63)$$

We can also rescale $V_Q(x)$ such that there is a single free parameter σ in Eq. (2.62) to yield

$$-\frac{\sigma^2}{2}\nabla^2\psi(x) + V_Q(x)\psi(x) = E\,\psi(x). \qquad (2.64)$$

Solving for $V_Q(x)$ we obtain

$$V_Q(x) = E + \frac{\sigma/2\nabla^2\psi(x)}{\psi(x)} = E - \frac{1}{2} + \frac{1}{2\sigma^2\psi(x)}\sum_i (x - x_i)^2 e^{-(x-x_i)^2/2\sigma^2}. \qquad (2.65)$$

To determine the value of $V_Q(x)$ uniquely we can require that $\min V_Q(x) = 0$, which makes

$$E = -\min \frac{\sigma/2\nabla^2\psi(x)}{\psi(x)}$$

and $0 \le E \le 1/2$. Note that $\psi(x)$ is the eigenfunction of H and E is the lowest eigenvalue of the operator, which corresponds to the ground state. Given the data set, we expect $V_Q(x)$ to increase quadratically outside the data region and to exhibit local minima associated with the locations of highest sample density (clusters). This can be interpreted as clustering because the potential function attracts the data distribution function $\psi(x)$ to its minima, whereas the Laplacian drives it away, producing a complicated potential function in the space. We should remark that in this framework E sets the scale at which the minima are observed. This derivation can be easily extended to multidimensional data.

We can see that $V_Q(x)$ in Eq. (2.65) is also a "quantum" potential function that differs from $V(x)$ in Eq. (2.57) because it is associated with a quantum description of the information potential. For Gaussian kernels the two fields are similar to each other within the regions with samples and because the derivative of the Gaussian is a Gaussian, but Eq. (2.65) may present advantages because of the intrinsic normalization produced by the eigenvalue that may simplify the search scale for minima of the potential field. We can also estimate the information quantum forces as presented above in Eq. (2.58).

2.7 Extension to α-Information Potential with Arbitrary Kernels

Recall the definition of Renyi's α entropy given in Eq. (2.18). Thus, its nonparametric estimator with arbitrary kernel $\kappa_\sigma(x)$ with bandwidth σ is given by

$$\hat{V}_\alpha(X) = \frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(x_j - x_i)\right)^{\alpha-1}, \qquad (2.66)$$

which can be written as a sum of contributions from each sample x_j, denoted $\hat{V}_\alpha(x_j)$,

$$\begin{aligned}\hat{V}_\alpha(x_j) &\overset{\Delta}{=} \frac{1}{N^{\alpha-1}}\left(\sum_{i=1}^{N}\kappa_\sigma(x_j-x_i)\right)^{\alpha-1}\\ \hat{V}_\alpha(X) &= \frac{1}{N}\sum_{j=1}^{N}\hat{V}_\alpha(x_j)\end{aligned} \tag{2.67}$$

for all positive $\alpha \neq 1$. Note that this α information potential can be written as a function of $\hat{V}_2(x_j)$ as

$$\hat{V}_\alpha(x_j) = \frac{1}{N^{\alpha-2}}\left(\sum_{i=1}^{N}\kappa_\sigma(x_j-x_i)\right)^{\alpha-2}\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x_j-x_i) = \hat{p}^{\alpha-2}(x_j)\hat{V}_2(x_j), \tag{2.68}$$

which means that all the integer IP$_\alpha$ can be derived conceptually from the quadratic information potential by scaling them by the estimated (α-2) PDF at the point. Naturally (in analogy with physical potentials), we determine the α information forces by simply taking the derivative of these information potentials with respect to the particle location (sample value).

$$\begin{aligned}\hat{F}_\alpha(x_j) &\overset{\Delta}{=} \frac{\partial}{\partial x_j}\hat{V}_\alpha(x_j) = \frac{\alpha-1}{N^{\alpha-1}}\left(\sum_{i=1}^{N}\kappa_\sigma(x_j-x_i)\right)^{\alpha-2}\left(\sum_{i=1}^{N}\kappa'_\sigma(x_j-x_i)\right)\\ &= (\alpha-1)\hat{p}_X^{\alpha-2}(x_j)\hat{F}_2(x_j).\end{aligned} \tag{2.69}$$

This formula defines the total information force acting on sample x_j, where the quadratic information force is similar to Eq. (2.59), with the exception that the kernel function need not be specifically Gaussian. In Eq. (2.68), the quadratic force is defined as

$$\hat{F}_2(x_j) \overset{\Delta}{=} \frac{1}{N}\left(\sum_{i=1}^{N}\kappa'_\sigma(x_j-x_i)\right). \tag{2.70}$$

From Eq. (2.69), which is the total information force acting on sample x_j, and using the additivity of quadratic forces in Eq. (2.70), we can write out the individual contributions of every other sample as

$$\hat{F}_\alpha(x_j;x_i) = (\alpha-1)\hat{p}_X^{\alpha-2}(x_j)\hat{F}_2(x_j;x_i), \tag{2.71}$$

where we defined

$$\hat{F}_2(x_j;x_i) \overset{\Delta}{=} \kappa'_\sigma(x_j-x_i). \tag{2.72}$$

Although we considered above only the single-dimensional case, extensions of these information potential and information force definitions to multidimensional situations is trivial. Note that, in choosing multidimensional kernel functions, some restrictions apply as mentioned in Section 2.3.

Notice that the generalized information forces introduce a scaling factor that depends on the estimated probability density of the corresponding sample and the selected entropy order. Specifically, the baseline is obtained for α = 2; that is, the quadratic information potential treats equally the contributions of all the samples. For

$\alpha > 2$, the scale factor (power of the estimated PDF) in Eq. (2.69) becomes a monotonically increasing function of the PDF value, meaning that compared to the quadratic case, the forces experienced by samples in dense regions of the sample space are amplified. For $\alpha < 2$, on the other hand, the opposite takes place, and the forces on sparse regions of the data space are amplified.

This scenario also shows the difficulty of estimating Shannon entropy directly from samples with kernel estimators. The information potential field estimated by Eq. (2.67) becomes constant over the space of the samples for α = 1, therefore from Eq. (2.69) the force becomes zero. This does not mean that for Shannon's entropy the individual interactions of samples are constant and their forces are zero everywhere in the space, but simply that Eqs. (2.67) and (2.69) that capture macroscopic behavior cannot be applied for α = 1. Renyi's entropy is discontinuous at α = 1, therefore the direct substitution of this value in the expressions should be avoided. However, we can use the above estimator formalism for values of alpha close to 1, but we can expect very slow convergence.

2.8. Renyi's Divergence and Mutual Information

The structure of probability spaces is much more complex than linear spaces, therefore computing distances in such spaces is nontrivial. The most widely used disimilarity measure is the Kullback-Leibler divergence [188] due to its nice properties (invariant to reparameterization, monotonicity for Markov chains, and linked locally to the Fisher information matrix that quantifies the Riemannian metric of the space) as we have briefly outlined in Chapter 1. In this section we present Renyi's definition of divergence and mutual information and will also propose two alternate dissimilarity measures in probability spaces that can be easily estimated nonparametrically with the information potential.

Renyi's α Divergence

Alfred Renyi, in his studies of information theory [264], proposed what is now called the *Renyi's divergence*, intrinsically linked with his definition of entropy and an extension to the KL divergence. The definition of this divergence measure and some of its basic properties are reviewed herein.

Renyi's order-α divergence of $g(x)$ from $f(x)$ is defined as [264]

$$D_\alpha(f \,||\, g) \overset{\Delta}{=} \frac{1}{\alpha - 1} \log \int_{-\infty}^{\infty} f(x) \left(\frac{f(x)}{g(x)} \right)^{\alpha - 1} dx. \qquad (2.73)$$

Property 2.10: Renyi's divergence measure has the following properties

i. $D_\alpha(f \,||\, g) \geq 0, \ \forall f, g, \ \alpha > 0.$

ii. $D_\alpha(f \,||\, g) = 0$ iff $f(x) = g(x) \ \forall x \in R.$

iii. $\lim_{\alpha \to 1} D_\alpha(f \,||\, g) = D_{KL}(f \,||\, g).$

Proof. We do the proof of each part separately.

i. Using Jensen's inequality on the argument of the logarithm in Eq. (2.73), we get

$$\int_{-\infty}^{\infty} f(x)\left(\frac{g(x)}{f(x)}\right)^{1-\alpha} dx \underset{0<\alpha<1}{\overset{\alpha>1}{\underset{\le}{\ge}}} \left(\int_{-\infty}^{\infty} f(x)\left(\frac{g(x)}{f(x)}\right)dx\right)^{1-\alpha} = 1. \tag{2.74}$$

Substituting this result in Eq. (2.73), the desired inequality for all $\alpha > 0$ is obtained.

ii. Clearly, if $g(x) = f(x)$, then $D_\alpha(f \parallel g) = 0$. For the reverse direction, suppose we are given that $D_\alpha(f \parallel g) = 0$. Assume $g(x) \neq f(x)$, so that $g(x) = f(x) + \delta(x)$, where $\int_{-\infty}^{\infty} \delta(x) = 0$, and $\exists x \in R$ such that $\delta(x) \neq 0$. Consider the divergence between these two PDFs. Equating this divergence to zero, we obtain

$$D_\alpha(f \parallel g) = \frac{1}{1-\alpha}\log \int_{-\infty}^{\infty} f(x)\left(\frac{f(x)+\delta(x)}{f(x)}\right)^{1-\alpha} dx = \frac{1}{1-\alpha}\log \int_{-\infty}^{\infty} f(x)\left(1+\frac{\delta(x)}{f(x)}\right)^{1-\alpha} dx = 0 \tag{2.75}$$

which implies that

$$\int_{-\infty}^{\infty} f(x)\left(1+\frac{\delta(x)}{f(x)}\right)^{1-\alpha} dx = 1 \;\Rightarrow\; \left(1+\frac{\delta(x)}{f(x)}\right) = 1, \qquad \forall x \in R. \tag{2.76}$$

From this last result, we get that $\delta(x) = 0,\ \forall x \in R$, which contradicts our initial assumption, therefore, we conclude that $g(x) = f(x)$.

iii. Consider the limit of Eq. (2.73) as $\alpha \to 1$.

$$\begin{aligned}
\lim_{\alpha\to 1} D_\alpha(f \parallel g) &= \lim_{\alpha\to 1} \frac{1}{\alpha-1}\log \int_{-\infty}^{\infty} f(x)\left(\frac{f(x)}{g(x)}\right)^{\alpha-1} dx \\
&= \frac{\lim\limits_{\alpha\to 1} \int_{\infty-}^{\infty} - f(x)\left(\frac{f(x)}{g(x)}\right)^{\alpha-1} \log\left(\frac{g(x)}{f(x)}\right)dx}{\lim\limits_{\alpha\to 1} \int_{\infty-}^{\infty}\left(\frac{f(x)}{g(x)}\right)^{\alpha-1} dx} \\
&= \int_{\infty-}^{\infty} f(x)\log\left(\frac{f(x)}{g(x)}\right)dx = D_{KL}(f \parallel g).
\end{aligned} \tag{2.77}$$

Following the same ideas used in deriving the estimator for Renyi's entropy, we can determine a kernel-based resubstitution estimate of Renyi's order-α divergence using Eq. (2.18). Suppose we have the i.i.d. samples $\{x_g(1),..., x_g(N)\}$ and $\{x_f(1),..., x_f(N)\}$ drawn from $g(x)$ and $f(x)$, respectively. The nonparametric estimator for Renyi's divergence obtained with this approach is given as

$$D_\alpha(f\|g)=\frac{1}{\alpha-1}\log E_p\left[\left(\frac{f(x)}{g(x)}\right)^{\alpha-1}\right]\approx\frac{1}{\alpha-1}\log\frac{1}{N}\sum_{j=1}^{N}\left(\frac{\hat{f}(x(j))}{\hat{g}(x(j))}\right)^{\alpha-1}$$

$$=\frac{1}{\alpha-1}\log\frac{1}{N}\sum_{j=1}^{N}\left(\frac{\sum_{i=1}^{N}\kappa_\sigma(x_f(j)-x_f(i))}{\sum_{i=1}^{M}\kappa_\sigma(x_g(j)-x_g(i))}\right)^{\alpha-1}=\hat{D}_\alpha(f\|g), \tag{2.78}$$

with the computational complexity $O(N^2)$, the same as the entropy estimator.

Although Renyi's α-divergence has many of the same properties as KL divergence, it is not as general. In fact, for Shannon's relative entropy, the total information gained by observing a random event *A* with probability *f*(*x*) that changes to *g*(*x*) by observing a second event *B* can be computed either by averaging the partial gains of information, or by averaging the increase in uncertainty with a negative sign [65], as we have seen in Section 1.5. Shannon information gain for continuous variables

$$I(f\|g)=\int f(x)\log\frac{f(x)}{g(x)}dx \tag{2.79}$$

is obtained by averaging over *f*(*x*) the partial gains of information log(*f*(*x*)/*g*(*x*)). Notice that Eq. (2.79) is equivalent to Shannon's relative entropy (KL divergence). However, if Renyi's entropy with $\alpha\neq 1$ is used to evaluate this gain of information or the negated increase in uncertainty the results differ. Renyi's gain of information by partial increase in information of order α is [266]

$$\bar{I}_\alpha(f\|g)=\frac{1}{1-\alpha}\log\int\frac{f(x)^{2-\alpha}}{g(x)^{1-\alpha}}dx. \tag{2.80}$$

If one uses the measure of uncertainty of order α, we get Renyi's gain of information of order-α or Renyi's divergence of Eq. (2.73) which is a different quantity (in fact $\bar{I}_\alpha(f\|g)=D_{2-\alpha}(g\|f)$). Therefore, Shannon relative entropy is the only one for which the sum of average gain of information is equal to the negated average increase of uncertainty. This different behavior between Shannon and Renyi stems from the generalized additivity used in Renyi's definition which excludes the case $I(f_\Gamma)+I(f_{-\Gamma})=0$, unless I_k are all the same (uniform distribution), where –Γ is the set of the negated amount of information (i.e., $-I_k$). This has implications for the definition of Renyi's mutual information as well.

Renyi's α Mutual Information

Recall that Shannon's mutual information between the components of an *n*-dimensional random vector *X* is equal to the KL divergence of the joint distribution of *X* from the product of the marginal distributions of the components of *X* [266]. Similarly, Renyi's α mutual information is defined as Renyi's divergence between the same quantities. Letting $p_X(.)$ be the joint distribution and $p_{X_o}(.)$ be the marginal density of the *o*th component, Renyi's mutual information becomes [266]

$$I_\alpha(X) \overset{\Delta}{=} \frac{1}{\alpha-1}\log \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{p_X^\alpha(x_1,\ldots,x_n)}{\prod_{o=1}^{n} p_{X_0}^{\alpha-1}(x_o)}\, dx_1..dx_n. \tag{2.81}$$

Once again, it is possible to write a kernel-based estimator for Renyi's mutual information by approximating the joint expectation with a sample mean

$$I_\alpha(X) \overset{\Delta}{=} \frac{1}{\alpha-1}\log E_X\left[\left(\frac{p_X(x_1,\ldots,x_n)}{\prod_{o=1}^{n} p_{X_o}(x_o)}\right)^{\alpha-1}\right] \approx \frac{1}{\alpha-1}\log\frac{1}{N}\sum_{j=1}^{N}\left(\frac{p_X(x(j))}{\prod_{o=1}^{n} p_{X_o}(x_o(j))}\right)^{\alpha-1} \tag{2.82}$$

and when replacing the PDFs with their Parzen estimators that use consistent kernels between the marginal and joint PDF estimates as mentioned in Property 2.7, the nonparametric mutual information estimator becomes

$$\hat{I}_\alpha(X) \overset{\Delta}{=} \frac{1}{\alpha-1}\log\frac{1}{N}\sum_{j=1}^{N}\left(\frac{\left(\frac{1}{N}\sum_{i=1}^{N}\kappa_\Sigma(x(j)-x(i))\right)}{\prod_{o=1}^{n}\left(\frac{1}{N}\sum_{i=1}^{N}\kappa_{\sigma_o}(x_o(j)-x_o(i))\right)}\right)^{\alpha-1} = \frac{1}{\alpha-1}\log\frac{1}{N}\sum_{j=1}^{N}\left(\frac{\left(\frac{1}{N}\sum_{i=1}^{N}\prod_{o=1}^{n}\kappa_{\sigma_o}(x(j)-x(i))\right)}{\prod_{o=1}^{n}\left(\frac{1}{N}\sum_{i=1}^{N}\kappa_{\sigma_o}(x_o(j)-x_o(i))\right)}\right)^{\alpha-1} \tag{2.83}$$

The limit of Eq. (2.83) when α → 1 is an estimate of Shannon's mutual information between the random variables under consideration. Therefore, this nonparametric mutual information estimator can be used to estimate directly Renyi's mutual information $I_\alpha(X)$ from data for α close to one, but it does not have all the nice properties of Shannon mutual information. Although it is nonnegative and symmetric, it may yield a value greater than 1 (i.e., the information on x_1 given by x_2 can be larger than the information of x_1, which is a shortcoming). There are many other alternatives to define Renyi's mutual information and unfortunately all of them have shortcomings. See Renyi [266] for a full treatment.

2.9. Quadratic Divergences and Mutual Information

As pointed out by Kapur [177], there is no reason to restrict ourselves to Shannon's measure of entropy or to Kullback-Leibler's measure for cross-entropy (density dissimilarity). Entropy and relative entropy are too deep and too complex concepts to be measured by a single measure under all conditions. This section defines divergence and mutual information measures involving only a simple quadratic form of PDFs to take direct advantage of the IP and its nice estimator. A geometric approach is used here.

Looking at the Euclidean space, the two most common families of distance are the sums of difference squares in coordinates and the inner-product distances, and they are be the starting point to derive the corresponding divergences in probability spaces. Equal neighborhoods in the simplex are transformed into spheres of equal size to preserve the Fisher information matrix [32]. This gives rise to a unit sphere where each transformed PMF has coordinates $\sqrt{p_k}$ (the simplex is transformed to the positive hyperoctant of the sphere). The geodesic distance D_G between two PMFs f and g in the sphere (i.e., the length of the great circle) can be estimated by the cosine of the angle between them, or $\cos D_G = \sum_k \sqrt{f_k}\sqrt{g_k}$. This result is related to the argument of the Bhattacharyya distance [39], which is defined (for continous PDFs) as $D_B(f,g)$

$$D_B(f,g) = -\ln\left(\int \sqrt{f(x)g(x)}dx\right) \tag{2.84}$$

$D_B(f, g)$ vanishes iff $f = g$ almost everywhere. We can further establish a link of Eq. (2.84) with Renyi's divergence with $\alpha = 1/2$ (apart from a scalar). The Chernoff distance [56] or generalized Bhattacharya distance is a non-symmetric measure defined by

$$D_C(f,g) = -\ln\left(\int (f(x))^{1-s}(g(x))^s dx\right) \qquad 0 < s < 1 \tag{2.85}$$

which for s =1/2 yields the Bhattacharyya, and again, apart from the scaling, corresponds to Renyi's divergence for $\alpha = 1 - s$.

Instead of the inner product distance we can also measure the distance between f and g in a linear projection space of the hyperoctant (chordal distance) as $D_H = \left(\sum_k (\sqrt{f_k} - \sqrt{g_k})^2\right)^{1/2}$. This result yields the Hellinger's distance [19] which is defined (for continuous densities) as

$$D_H(f,g) = \left[\int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx\right]^{1/2} = \left[2\left(1 - \int \sqrt{f(x)g(x)}dx\right)\right]^{1/2}. \tag{2.86}$$

Compared with the KL and Renyi's divergences, Hellinger's distance has the advantage of being a difference of PDFs so it avoids stability problems when the denominator PDF is zero. It is also related to the Havrda-Charvat (α = 1/2) divergence that is intimately related to α Renyi's divergence.

After all, Bhattacharyya Eq. (2.84) and Hellinger's distances Eq. (2.86) are angular and Euclidean distances in the hypersphere and their relationship with Renyi's α = 1/2 divergence inspired us to seek definitions of divergences that could benefit from the α information potential estimator (specifically α = 2).

The similarity between two PDFs using the 2-norm is a simple and straightforward distance measure; it obeys all the properties of a distance (including symmetry and the triangular inequality), and can be written as $D_{ED}(f,g) = \int \sqrt{(f(x) - g(x))^2} dx$. For simplicity we do not include the square root in the definition because our goal is to use these measures as cost functions, so the Euclidean distance between PDFs is redefined as

$$D_{ED}(f,g) = \int (f(x) - g(x))^2 dx = \int f^2(x)dx - 2\int f(x)g(x)dx + \int g^2(x)dx \tag{2.87}$$

$D_{ED}(f, g)$ can be recognized as belonging to the same family as the Herlinger distance (chordal distances) but with a different α-norm. Although being a distance, $D_{ED}(f, g)$ is sometimes lumped with the divergence terminology in PDF spaces.

The squared distance between the joint PDF and the factorized marginal PDF is called the *quadratic mutual information Euclidean distance* (QMI$_{ED}$), and is written as

$$I_{ED}(X_1, X_2) = D_{ED}(f_{X_1X_2}(x_1, x_2),\ f_{X_1}(x_1) f_{X_2}(x_2)). \tag{2.88}$$

$D_{ED}(f,g) \geq 0$ with equality if and only if $f(x) = g(x)$ almost everywhere and the integrals involved are all quadratic forms of PDFs. Obviously, the QMI$_{ED}$ between X_1 and X_2 is nonnegative and is zero if and only if $f_{X_1X_2}(x_1,x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$; that is, X_1

and X_2 are independent random variables. There is no strict theoretical justification that the QMI$_{ED}$ is an appropriate measure for dependence between two variables. However, it can be shown that $D_{ED}(f, g)$ is a lower bound for the KL divergence [319], therefore when one maximizes $D_{ED}(f, g)$, we are also maximizing KL. For multiple variables, the extension of QMI$_{ED}$ interpreted as a multivariate dissimilarity measure is straightforward:

$$I_{ED}(X_1,..., X_k) = D_{ED}\left(f_X(x_1,..., x_k), \prod_{i=1}^{k} f_{X_i}(x_i)\right),$$

where $f_X(x_1,..., x_k)$ is the joint PDF, and $f_{X_i}(x_i), (i = 1,..., k)$ are marginal PDFs.

The other possible PDF divergence is related to the Battacharyya distance. Formally it can be derived from the Cauchy-Schwarz inequality [276]:

$$\sqrt{\int f^2(x)dx \int g^2(x)dx} \geq \int f(x)g(x)dx, \tag{2.89}$$

where equality holds if and only if $f(x) = cg(x)$ for a constant scalar c. If $f(x)$ and $g(x)$ are PDFs (i.e., $\int f(x)dx = 1$ and $\int g(x)dx = 1$), then $f(x) = cg(x)$ implies c = 1. So, for two PDFs $f(x)$ and $g(x)$ equality holds if and only if $f(x) = g(x)$. Similarly to $D_{ED}(f, g)$ for the estimators we normally use the square of Eq. (2.89) to simplify the calculations. Thus, we may define the Cauchy- Schwarz divergence for two PDFs as

$$D_{CS}(f, g) = -\log \frac{\left(\int f(x)g(x)dx\right)^2}{\int f^2(x)dx \int g^2(x)dx}, \tag{2.90}$$

$D_{CS}(f, g) \geq 0$, where the equality holds if and only if $f(x) = g(x)$ almost everywhere and the integrals involved are all quadratic forms of PDFs. $D_{CS}(f, g)$ is symmetric but it does not obey the triangular inequality.

Let us look closely at Eq. (2.87). One can immediately recognize the first and last terms as the quadratic information potential of $f(x)$ and $g(x)$, respectively. The middle term $\int f(x)g(x)dx$ is called the *cross information potential (CIP)*, and basically estimates the interactions on locations in the space dictated by the dataset $f(x)$ in the potential created by the dataset $g(x)$ (or viceversa). This is really the term that measures the "distance" between the two PDFs, because the other two are simply normalizing terms. The $D_{CS}(f, g)$ of Eq. (2.90) can be rewritten as

$$D_{CS}(f, g) = \log \int f(x)^2 dx + \log \int g(x)^2 dx - 2\log \int f(x)g(x)dx, \tag{2.91}$$

where all the three terms of $D_{ED}(f, g)$ appear also in $D_{CS}(f, g)$, simply with a logarithmic weighting. Based on $D_{CS}(f, g)$, we define the *Cauchy–Schwarz quadratic mutual information* (QMI$_{CS}$) between two variables X_1 and X_2 as

$$I_{CS}(X_1, X_2) = D_{CS}\left(f_X(x_1, x_2), f_{X_1}(x_1) f_{X_2}(x_2)\right) , \tag{2.92}$$

where the notations are the same as above. Directly from above, $I_{CS}(X_1, X_2) \geq 0$ meets the equality if and only if X_1 and X_2 are independent random variables. So, I_{CS} is an appropriate measure of independence. This measure is a geodesic distance in the sphere, therefore the Cauchy–Schwarz divergence may also be appropriate as a dependence measure in cases where the PDFs exist in the sphere. For multivariate variables, the extension of QMI$_{CS}$ is also straightforward:

$$I_{CS}(X_1,\ldots,X_k)=D_{CS}\left(f_X(x_1,\ldots,x_k),\prod_{i=1}^{k}f_{X_i}(x_i)\right).$$

Cauchy-Schwarz Divergence and Renyi's Relative Entropy

Recently, Lutwak, et al [205] defined a new Renyi's divergence called the relative α-Renyi entropy between *f*(*x*) and g(*x*) as

$$D_{R_\alpha}(f,g)=\log\frac{\left(\int_R g^{\alpha-1}(x)f(x)\right)^{\frac{1}{(1-\alpha)}}\left(\int_R g^{\alpha}(x)\right)^{1/\alpha}}{\left(\int_R f^{\alpha}(x)\right)^{\frac{1}{(\alpha(1-\alpha))}}}. \tag{2.93}$$

Note that the denominator in the argument of the log now contains an integral that is more robust than Renyi's original definition of Eq. (2.73). So *f*(*x*) could be zero at some points of the domain but overall the integral is well defined, thus avoiding numerical issues of Eq. (2.73). Again, for α → 1, this gives $D_{KL}(f\,\|\,g)$. In particular, for α = 2 Eq. (2.93) is exactly the Cauchy–Schwarz divergence of Eq. (2.90). This is a very interesting relation because it provides an information-theoretic interpretation both for the Cauchy-Schwarz divergence and also for the integral of the product of PDFs. Indeed, we can rewrite the Cauchy–Schwarz divergence in terms of Renyi's quadratic entropy as

$$\begin{aligned}D_{CS}(X,Y)&=-2\log(\int f(x)g(x))^2dx+\log(\int f(x)^2dx)+\log(\int g(x)^2dx)\\&=2H_2(X;Y)-H_2(X)-H_2(Y),\end{aligned} \tag{2.94}$$

where the first term can be shown to be the *quadratic Renyi's cross-entropy* [259] (and should not be confused with the joint entropy of *X* and *Y*). The similarity of this expression with Shannon's mutual information in Eq. (1.10) is striking if we think in terms of cross-entropy versus joint entropy.

2.10. Information Potentials and Forces in the Joint Space

The interactions among the samples interpreted as information particles for the case of divergence are substantially more involved than IP because of the different information potential fields that exist. In essence one has to realize that each probability density function creates its own information potential field, and that particle interactions are produced by weighted sums of each potential field computed in the joint space [340]. We illustrate the principles for the calculation of the Euclidean distance and QMI$_{ED}$.

Euclidean and Cauchy–Schwarz Divergence Estimators

The divergences are composed of three different information potential fields, each specified by the location of the samples from *f*(*x*), from *g*(*x*), and the cross–information potential field. Because the potentials are additive, we can compute one at a time and add the result as specified by Eq. (2.86). The information potential estimated by the Gaussian kernel for each PDF is given by

$$\begin{aligned}\hat{V}_f &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i=j}^{N} G_{\sqrt{2}\sigma}(x_f(i)-x_f(j))^2 \\ \hat{V}_g &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i=j}^{N} G_{\sqrt{2}\sigma}(x_g(i)-x_g(j))^2 \\ \hat{V}_c &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i=j}^{N} G_{\sqrt{2}\sigma}(x_f(i)-x_g(j))^2,\end{aligned} \tag{2.95}$$

where for simplicity we assume that we have the same number of samples (N) in each dataset and the same kernel size σ. V_C is the cross-information potential estimator and basically measures the interaction of the field created by $f(x)$ on the locations specified by $g(x)$. The Euclidean and Cauchy–Schwarz information potential fields are therefore:

$$\begin{aligned}\hat{D}_{ED}(f,g) &= \hat{V}_{ED} = \hat{V}_f + \hat{V}_g - 2\hat{V}_c \\ \hat{D}_{CS}(f,g) &= \hat{V}_{CS} = \log\frac{\hat{V}_f\hat{V}_g}{\hat{V}_c^{\,2}}.\end{aligned} \tag{2.96}$$

The computation of the Euclidean and Cauchy–Schwarz information forces exerted on each sample x_i can be easily achieved using the additive rule of derivatives

$$\begin{aligned}\frac{\partial\hat{V}_{ED}}{\partial x_i} &= \frac{\partial\hat{V}_f}{\partial x_i} + \frac{\partial\hat{V}_g}{\partial x_i} - 2\frac{\partial\hat{V}_c}{\partial x_i} \\ \frac{\partial\hat{V}_{CS}}{\partial x_i} &= \frac{1}{\hat{V}_f}\frac{\partial\hat{V}_f}{\partial x_i} + \frac{1}{\hat{V}_g}\frac{\partial\hat{V}_g}{\partial x_i} - \frac{2}{\hat{V}_c}\frac{\partial\hat{V}_c}{\partial x_i}.\end{aligned} \tag{2.97}$$

The computation of the Euclidean and Cauchy–Schwarz divergences can proceed in a fashion very similar to the information potential defining a matrix of distances and of scalars as given by Eq. (2.61). See [340] for a full treatment.

Generalized Information Potential (GIP) for Quadratic Mutual Information

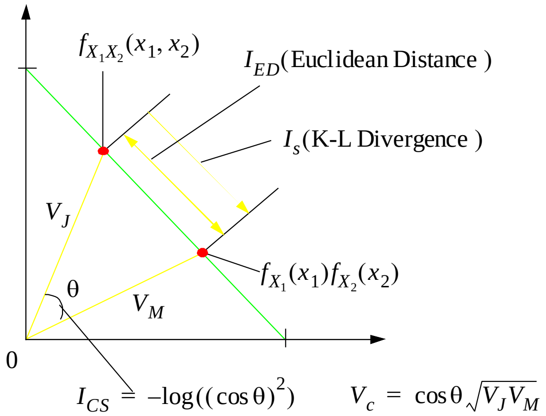
Both QMI_{ED} and QMI_{CS} can be written in a very similar way to $D_{ED}(f, g)$ and $D_{CS}(f, g)$ but they are a little more complex due to the fact that we have two variables and the existence of the joint and the product of the marginals. We can decompose the overall expressions Eqs. (2.88) and (2.92) in the following three terms,

$$\begin{cases} V_J = \iint f_{X_1X_2}(x_1, x_2)^2 dx_1 dx_2 \\ V_M = \iint (f_{X_1}(x_1) f_{X_2}(x_2))^2 dx_1 dx_2 \\ V_c = \iint f_{X_1X_2}(x_1, x_2) f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 \end{cases} \tag{2.98}$$

where V_J is the IP of the joint PDF, V_M is the IP of the factorized marginal PDF, and V_C is the generalized cross information potential. Just like for the quadratic divergences, this is really the term that measures the interactions between the two information potentials, whereas the other two are proper normalizations. With these three terms, both QMIs yield

$$\begin{cases} I_{ED} = V_J - 2V_c + V_M \\ I_{CS} = \log V_J - 2\log V_c + \log V_M \end{cases} \tag{2.99}$$

Figure 2.8 shows the illustration of the geometrical interpretation of all these quantities in the 2D simplex (for the case of discrete random variables).



$f_{X_1X_2}(x_1, x_2)$
I_{ED}(Euclidean Distance)
I_s(K-L Divergence)
V_J
θ
V_M
$f_{X_1}(x_1)f_{X_2}(x_2)$
0
$I_{CS} = -\log((\cos\theta)^2)$
$V_c = \cos\theta\sqrt{V_J V_M}$

Figure 2.8. Geometrical interpretation of quadratic mutual information.

I_S, as previously mentioned, is the KL divergence between the joint PDF and the factorized marginal PDF, I_{ED} is the squared Euclidean distance between these two PDFs, and I_{CS} is related to the angle between these two PDFs.

For the estimation of each of the potentials in Eq. (2.98) the following notation is be used: subscripts denote the input components, and indices represent sums over samples. For a given dataset $\{\boldsymbol{x}(i) = (x_1(i), x_2(i))^T \mid i = 1,..., N\}$ of a two-dimensional variable $X = (x_1, x_2)^T$, the joint and marginal PDFs define a joint (V_J), a marginal (V_M) and a cross (V_C) information potential field from Eq. (2.98). Using Gaussian kernels to estimate the joint and the marginals yields,

$$\begin{cases} \hat{f}_{X_1X_2}(x_1, x_2) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} G_\sigma(\boldsymbol{x} - \boldsymbol{x}(i)) \\ \hat{f}_{X_1}(x_1) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} G_\sigma(x_1 - x_1(i)) \\ \hat{f}_{X_2}(x_2) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} G_\sigma(x_2 - x_2(i)). \end{cases} \tag{2.100}$$

Because information potential fields are additive, we can estimate independently the three terms in QMI_{ED} or QMI_{CS} of Eq. (2.99) based only on the given dataset.

Note that V_J, which exists over the joint space, can be decomposed for radially symmetric kernels in a product of interactions along each of the variables $G(\boldsymbol{x}_i - \boldsymbol{x}_j) = G(x_{1i} - x_{1j})G(x_{2i} - x_{2j})$, where $\boldsymbol{x} = (x_1, x_2)^T$.The generalized cross–information potential V_C of Eq. (2.98) is the potential that seems more difficult to compute, therefore it is illustrated here. Starting from the definition of the information potential, we obtain

$$\begin{aligned}
\hat{V}_C &= \int\int \hat{f}(x_1, x_2)\,\hat{f}(x_1)\,\hat{f}(x_2)dx_1dx_2 \\
&= \int\int \left[\frac{1}{N}\sum_{k=1}^{N} G_\sigma(x_1 - x_1(k))G_\sigma(x_2 - x_2(k))\right]\left[\frac{1}{N}\sum_{i=1}^{N} G_\sigma(x_1 - x_1(i))\right]\left[\frac{1}{N}\sum_{j=1}^{N} G_\sigma(x_2 - x_2(j))\right]dx_1dx_2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{N}\sum_{k=1}^{N}\int G_\sigma(x_1 - x_1(i))G_\sigma(x_1 - x_1(k))dx_1\int G_\sigma(x_2 - x_2(k))G_\sigma(x_2 - x_2(j))dx_2 \\
&= \frac{1}{N}\sum_{k=1}^{N}\left[\frac{1}{N}\sum_{i=1}^{N} G_{\sqrt{2}\sigma}(x_1(k) - x_1(i))\right]\left[\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_2(k) - x_2(j))\right] \qquad (2.101)
\end{aligned}$$

Notice that the GCIP for QMI requires $O(N^{3})$ computation. V_M can be further factorized as two marginal information potentials V_1 and V_2

$$\begin{aligned}
\hat{V}_M &= \iint \hat{f}^{2}_{X_1}(x_1)\,\hat{f}^{2}_{X_2}(x_2)dx_1dx_2 \\
\hat{V}_1 &= \iint \hat{f}^{2}_{X_1}(x_1)dx_1 \qquad (2.102)\\
\hat{V}_2 &= \iint \hat{f}^{2}_{X_2}(x_2)dx_2
\end{aligned}$$

Therefore the final expressions are

$$\begin{aligned}
&\hat{V}_k(i, j) = G_{\sqrt{2}\sigma}(x_k(i) - x_k(j)), \qquad \hat{V}_k(i) = \frac{1}{N}\sum_{j=1}^{N}\hat{V}_k(i, j),\ \hat{V}_k = \frac{1}{N}\sum_{i=1}^{N}\hat{V}_k(i),\ \ k = 1,2 \\
&\hat{V}_J = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{V}_1(i, j)\hat{V}_2(i, j) \qquad (2.103)\\
&\hat{V}_M = \hat{V}_1\hat{V}_2 \quad \text{with } \hat{V}_k = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{V}_k(i, j),\ \ k = 1,2 \\
&\hat{V}_C = \frac{1}{N}\sum_{i=1}^{N}\hat{V}_1(i)\hat{V}_2(i).
\end{aligned}$$

So, the estimated Euclidean Mutual information (QMI_{ED}) and the estimated Cauchy-Schwarz Mutual Information (QMI_{CS}) are given by

$$\begin{aligned}
&\hat{I}_{ED}(X_1, X_2) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{V}_1(i, j)\hat{V}_2(i, j) + \hat{V}_1\hat{V}_2 - \frac{1}{N}\sum_{i=1}^{N}\hat{V}_1(i)\hat{V}_2(i) \\
&\hat{I}_{CS}(X_1, X_2) = \log\frac{\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{V}_1(ij)\hat{V}_2(i, j)\right)\left(\hat{V}_1\hat{V}_2\right)}{\left(\frac{1}{N}\sum_{i=1}^{N}\hat{V}_1(i)\hat{V}_2(i)\right)^2}. \qquad (2.104)
\end{aligned}$$

From the above, we can see that both QMIs can be expressed as interactions between the marginal information potential fields at different levels: $V_1(i, j)V_2(i, j)$ is the level of the sample-to-sample interactions from each marginal (the joint field), $V_1(i)V_2(j)$ is the level of one full marginal field acting on a single sample (the GCIP), and V_1V_2 is the interaction between both marginal potential fields (product of marginals). $\hat{I}_{ED}$ is called the Euclidean generalized information potential (GIP$_{\text{ED}}$), and $\hat{I}_{CS}$ is the Cauchy-Schwartz generalized information potential (GIP$_{\text{CS}}$).

The quadratic mutual information and the corresponding cross information potential can be easily extended to the case with multiple variables (e.g., $X = (x_1,\ldots,x_k)^T$). In this case, we have similar IPs and marginal IPs as in Eq. (2.104). Then we have the QMI$_{\text{ED}}$ and QMI$_{\text{CS}}$ and their corresponding GIP$_{\text{ED}}$ and GIP$_{\text{CS}}$ as follows,

$$\begin{aligned}
\hat{I}_{ED}(X_1,\ldots,X_K) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\prod_{k=1}^{K} \hat{V}_k(i,j) - \frac{2}{N}\sum_{i=1}^{N}\prod_{k=1}^{K} \hat{V}_k(i) + \prod_{k=1}^{K} \hat{V}_k \\
\hat{I}_{CS}(X_1,\ldots,X_K) &= \log \frac{\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\prod_{k=1}^{K} \hat{V}_k(i,j)\right)\prod_{k=1}^{K} \hat{V}_k}{\left(\frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{K} \hat{V}_k(i)\right)^2}.
\end{aligned} \qquad (2.105)$$

Generalized Information Forces

Three different potentials contribute to the generalized information potential, but because the derivative is distributive with respect to addition, one can still operate on each term independently. The cases of GIP$_{\text{ED}}$ and GIP$_{\text{CS}}$ are slightly different because of the logarithm, but the procedure is to take the derivative of Eq. (2.104) with respect to a given sample, yielding

$$\begin{aligned}
\hat{F}_{ED}(i) &= \frac{\partial \hat{I}_{CS}}{\partial x_k(i)} = \frac{\partial \hat{V}_j}{\partial x_k(i)} - \frac{2\partial \hat{V}_C}{\partial x_k(i)} + \frac{\partial \hat{V}_k}{\partial x_k(i)} \\
\hat{F}_{CS}(i) &= \frac{\partial \hat{I}_{CS}}{\partial x_k(i)} = \frac{1}{V_j}\frac{\partial \hat{V}_j}{\partial x_k(i)} - \frac{2}{V_C}\frac{\partial \hat{V}_C}{\partial x_k(i)} + \frac{1}{V_k}\frac{\partial \hat{V}_k}{\partial x_k(i)}.
\end{aligned} \qquad (2.106)$$

The case for multiple variables can be readily obtained in a similar way (see [340] for a full treatment).

A simple example

To understand the similarities and differences among the dissimilarity measures I_S, I_{ED}, and I_{CS}, let's look at a simple case with two discrete random variables X_1 and X_2 as shown in Figure 2.9. It is trivial to apply these definitions to discrete events. We exemplify here the QMI$_{\text{ED}}$ and QMI$_{\text{CS}}$. For the discrete variables X_1 and X_2 with probability distribution $\{P_{X_1}(i); i=1,\ldots,n\}$ and $\{P_{X_2}(j); j=1,\ldots,m\}$, respectively, and the joint probability distribution $\{P_X(i,j); i=1,\ldots,n; j=1,\ldots,m\}$, the QMI$_{\text{ED}}$ and QMI$_{\text{CS}}$ are

$$\begin{cases} I_{ED}(X_1,X_2)=\sum_{i=1}^{n}\sum_{j=1}^{m}(P_X(i,j)-P_{X_1}(i)P_{X_2}(j))^2 \\ I_{CS}(X_1,X_2)=\log\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{m}(P_X(i,j))^2\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{m}(P_{X1}(i)P_{X_2}(j))^2\right)}{\sum_{i=1}^{n}\sum_{j=1}^{m}(P_X(i,j)P_{X1}(i)P_{X_2}(j))^2}. \end{cases} \quad (2.107)$$

X_1 can take the values 1 or 2 with a probability $P_{X_1}=(P_{X_1}(1),P_{X_1}(2))$; that is, $P(X_1=1)=P_{X_1}(1)$ and $P(X_1=2)=P_{X_1}(2)$. Similarly X_2 can take the values 1 or 2 with the probability $P_{X_2}=(P_{X_2}(1),P_{X_2}(2))$ where $P(X_2=1)=P_{X_2}(1)$ and $P(X_2=2)=P_{X_2}(2)$. The joint probability distribution is $P_X=(P_X(1,1),P_X(1,2),P_X(2,1),P_X(2,2))$; where $P_X(1,1)=P((X_1,X_2)=(1,1))$ and likewise for the other cases. Obviously, $P_{X_1}(1)=P_X(1,1)+P_X(1,2)$, $P_{X_1}(2)=P_X(2,1)+P_X(2,2)$, $P_{X_2}(1)=P_X(1,1)+P_X(2,1)$, and $P_{X_2}(2)=P_X(1,2)+P_X(2,2)$. In the following figures related to this example, the probability variables are simplified as $P_{X_1}^1=P_{X_1}(1), P_X^{11}=P_X(1,1)$, etc.

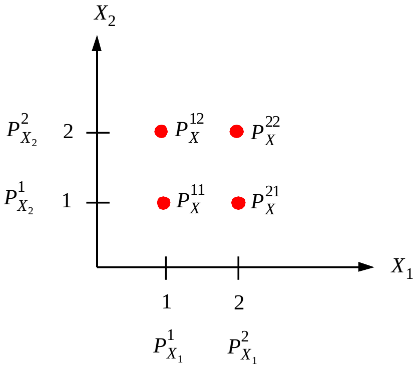


Figure 2.9. The 2D data for the example.

First, let's look at the case with the marginal distribution of X_1 fixed as P_{X1} = (0.6,0.4). Then the free parameters left are $P_X(1,1)$ from [0, 0.6] and $P_X(2,1)$ from [0, 0.4]. When $P_X(1,1)$ and $P_X(2,1)$ change in these ranges, the values of I_S , I_{ED} , and I_{CS} can be easily calculated. The right graphs in Figure 2.10 show the contour plots of the corresponding left surfaces (contour means that each line has the same functional value).

These graphs show that although the contours of the three measures are different, they reach the minimum value 0 in the same line $P_X(1,1) = 1.5P_X(2,1)$ where the joint probabilities equal the corresponding factorized marginal probabilities. And the maximum values, although different, are also reached at the same points ($P_X(1,1)$, $P_X(2,1)$) = (0.6, 0) and (0, 0.4) where the joint probabilities are

$$\begin{bmatrix} P_X(1,2) & P_X(2,2) \\ P_X(1,1) & P_X(2,1) \end{bmatrix}=\begin{bmatrix} 0 & 0.4 \\ 0.6 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} P_X(1,2) & P_X(2,2) \\ P_X(1,1) & P_X(2,1) \end{bmatrix}=\begin{bmatrix} 0.6 & 0 \\ 0 & 0.4 \end{bmatrix},$$

respectively.

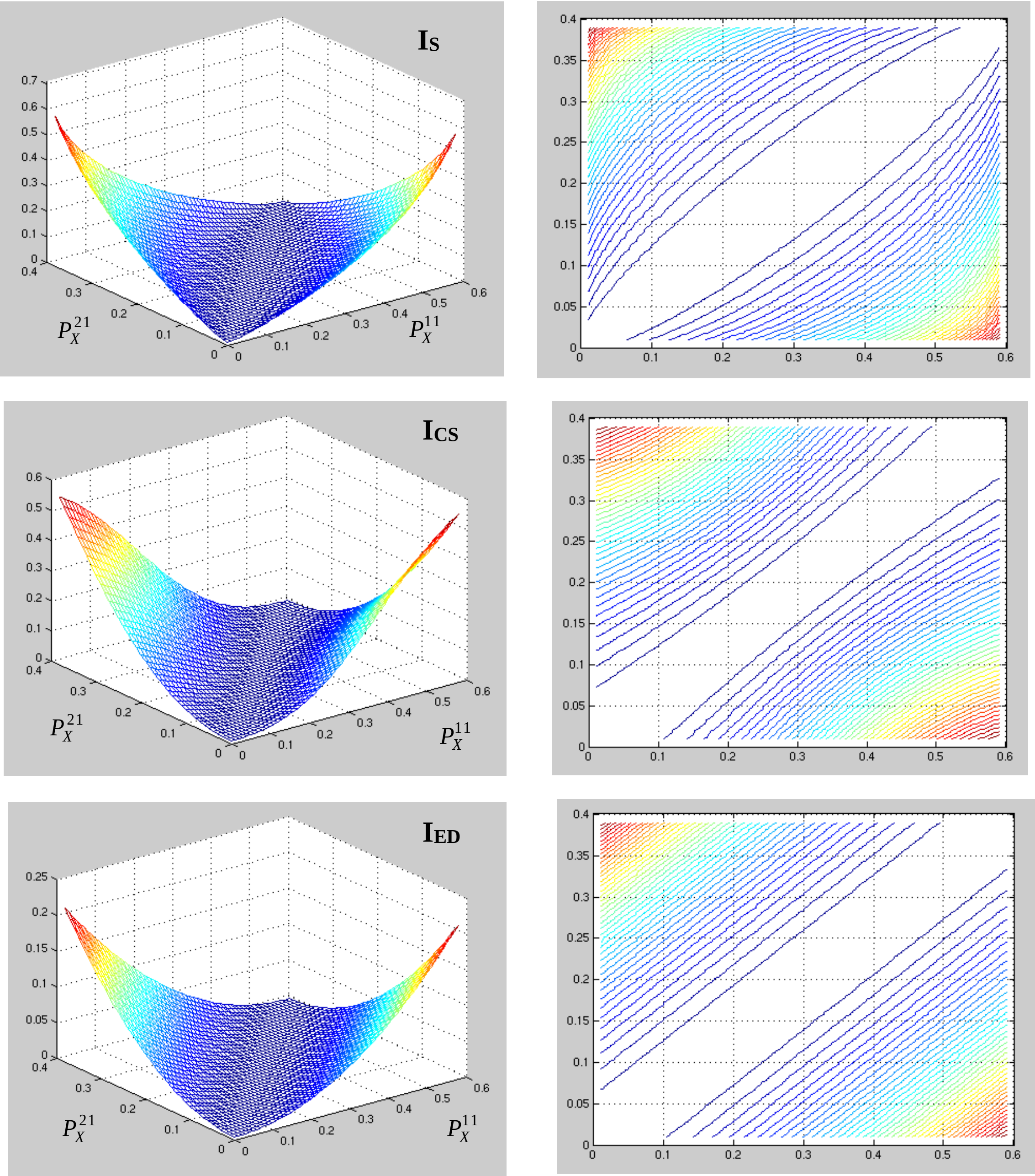


Figure 2.10. The surfaces and contours of I_S , I_{ED} , and I_{CS} versus $P_X(1,1)$ and $P_X(2,1)$

If the marginal probability of X_2 is further fixed (e.g. P_{X2} = (0.3,0.7)), then the free parameter is $P_X(1,1)$ from 0 to 0.3, which can be regarded as the previous setting with a further constraint specified by $P_X(1,1) + P_X(2,1) = 0.3$. In this case, both marginal probabilities of X_1 and X_2 are fixed, the factorized marginal probability distribution is also fixed and only the joint probability distribution will change. Figure 2.11 shows how the three measures change with $P_X(1,1)$, from which we can see that the minima are reached at the same point $P_X(1,1) = 0.18$, and the maxima are also reached at the same point $P_X(1,1) = 0$; that is,

$$\begin{bmatrix} P_X(1,2) & P_X(2,2) \\ P_X(1,1) & P_X(2,1) \end{bmatrix} = \begin{bmatrix} 0.6 & 0.1 \\ 0 & 0.3 \end{bmatrix}.$$

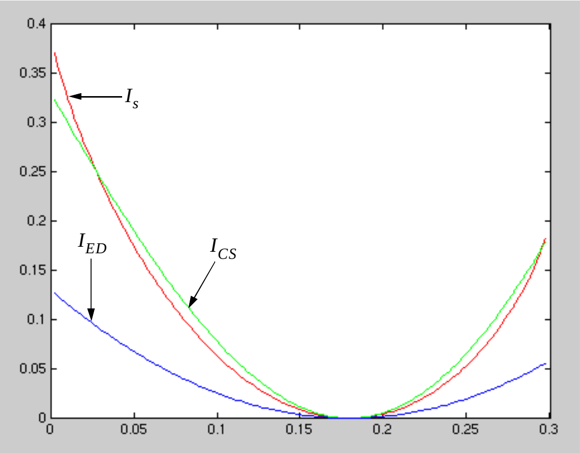


Figure 2.11. I_S , I_{ED} , and I_{CS} versus $P_X(1,1)$.

From this simple example, we can see that although the three measures are different, they have the same minimum point and also have the same maximum points in this particular case. It is known that both Shannon's mutual information I_S and QMI$_{ED}$ (I_{ED}) are convex functions of PDFs, and I_{ED} is a lower bound for I_S. From the above graphs, we can confirm this fact and also reach the conclusion that QMI$_{CS}$ (I_{CS}) is not a strictly convex function of PDFs.

2.11. Fast Computation of IP and CIP

One of the practical difficulties of the information potential, cross-information potential and ITL quantities in general, such as D_{ED} and D_{CS} and QMI$_{ED}$ and QMI$_{CS}$ is that the calculations are $O(N^2)$ or $O(N^3)$, respectively. This section presents an effort to make the estimation of IP faster using two techniques: one based on the fast Gauss transform (FGT) [122] and the other exploiting using the incomplete Cholesky decomposition the Gram matrix band structure that for kernels is known to possess rapidly decreasing eigenvalues, particularly in low dimensions [100].

Fast Gauss Transform

The fast multipole method is a very interesting and important family of fast evaluation algorithms that have been developed over the past two decades to enable rapid calculation of approximations, with arbitrary accuracy, to large matrix-vector products of the form **Ad** where the elements of **A** are $a_{i,j} = \sum_i \sum_j \varphi(x_i - x_j)$ with φ a nonlinear fast decaying positive function of the argument [121]. The fast Gauss transform [122] is a special case derived for efficient calculation of weighted sums of unidimensional Gaussians at a point y_i,

$$S(y_i) = \sum_{j=1}^{N} w_i e^{-(y_j - y_i)^2 / 4\sigma^2} \quad i = 1,..., M \qquad (2.108)$$

The FGT has been applied to many areas including astrophysics, kernel density estimation, and machine learning algorithms decreasing the computation from $O(NM)$ to $O(N + M)$ where N is the number of samples (sources) and M the number of points where the evaluation is required. The computational savings come from two facts, both related to the shifting property of the Gaussian function

$$e^{-\left(\frac{y_j - y_i}{\sigma}\right)^2} = e^{-\left(\frac{y_j - y_C - (y_i - y_C)}{\sigma}\right)^2} = e^{-\left(\frac{y_j - y_c}{\sigma}\right)^2} \sum\nolimits_{n=0}^{\infty} \frac{1}{n!}\left(\frac{y_i - y_c}{\sigma}\right)^n h_n\left(\frac{y_j - y_c}{\sigma}\right), \qquad (2.109)$$

which means that a Gaussian centered at y_j can be shifted to a sum of Hermite polynomials times a Gaussian, all centered at y_c. First, the Hermite polynomials $h_n(y)$ given by

$$h_n(y) = (-1)^n \frac{d^n \exp(-x^2)}{dx^n} \qquad (2.110)$$

are very efficient in the approximation and a small order p is normally sufficient; that is,

$$\exp\left(\frac{-(y_j - y_i)^2}{4\sigma^2}\right) = \sum_{n=0}^{p-1} \frac{1}{n!}\left(\frac{y_i - y_C}{2\sigma}\right)^n h_n\left(\frac{y_j - y_C}{2\sigma}\right) + \varepsilon(p),$$

where $\varepsilon(p)$ is the error associated with the truncation of the expansion at order p. The second savings is that there is no need to evaluate every Gaussian at every point. Instead a p-term sum is computed around a small number y_c of cluster centers with $O(Np)$ computation with Eq. (2.109). These sums are then shifted to the y_i desired locations and computed in another $O(Mp)$ operation. In practice, an expansion around a single center is not always accurate over the entire domain of interest. A tiling of the space is constructed and the Gaussian function is expanded at multiple centers with the FGT. To efficiently subdivide the space, a very simple greedy algorithm called *furthest-point clustering* [117] can be used, which computes a data partition with a maximum radius at most twice the optimum for the problem. The direct implementation of furthest-point clustering has running time $O(BN)$, with B the number of clusters.

If one recalls the definition of the IP this algorithm can be immediately applied, remembering that now the sources and the locations where the expansion needs to be computed coincide. If we apply this expansion to the IP $V(y)$, we obtain

$$V(y) \approx \frac{1}{2\sigma N^2 \sqrt{\pi}} \sum_{j=1}^{N} \sum_{b=1}^{B} \sum_{n=0}^{p-1} \frac{1}{n!} h_n\left(\frac{y_j - y_{C_b}}{2\sigma}\right) C_n(b), \qquad (2.111)$$

where B is the number of clusters used with centers y_{Cb} , and $C_n(b)$ is defined by

$$C_n(b) = \sum_{y_i \in B} \left(\frac{y_j - y_{C_b}}{2\sigma}\right)^n, \qquad (2.112)$$

From the above equation, we can see that the total number of operations required is $O(BpN)$ per data dimension. The order p of the expansion depends on the desired accuracy required (normally 4 or 5), and is independent of N. In addition to the complexity reduction, the other appeal of the FGT is that the code becomes parallelizable due to the clustering step.

Taylor Series for Multiple Dimensions

The extension to more than one dimension of the previous algorithm is done by treating the multivariate Gaussian as a Kronecker product of univariate Gaussians. Following the multi-index notation of the original FGT papers, we define the multidimensional Hermite function as

$$h_\alpha(\mathbf{y}) = h_{\alpha_1}(y_1)h_{\alpha_2}(y_2)\cdots h_{\alpha_d}(y_d), \tag{2.113}$$

where $\mathbf{y} = (y_1, \ \cdots, \ y_d)^T \in R^d$ and $\alpha = (\alpha_1, \ \ldots, \ \alpha_d) \in N^d$. As can be expected the algorithm scales up very poorly with dimension due to this product form.

An alternative method introduced by Yang et al. [345] is to expand the Gaussian function into a multivariate Taylor series. The Gaussian function is factorized as

$$\exp\left(-\frac{\|\mathbf{y}_j - \mathbf{y}_i\|^2}{4\sigma^2}\right) = \exp\left(-\frac{\|\mathbf{y}_j - c\|^2}{4\sigma^2}\right)\exp\left(-\frac{\|\mathbf{y}_i - c\|^2}{4\sigma^2}\right)\exp\left(2\frac{(\mathbf{y}_j - c)\cdot(\mathbf{y}_i - c)}{4\sigma^2}\right). \tag{2.114}$$

In the third term of (2.114), the product of the evaluation at two different points (called the entanglement) is split by expanding the exponential into a Taylor series as

$$\exp\left(2\frac{(\mathbf{y}_j - c)\cdot(\mathbf{y}_i - c)}{4\sigma^2}\right) = \sum_{\alpha\ge 0}\frac{2^{|\alpha|}}{\alpha!}\left(\frac{\mathbf{y}_j - c}{2\sigma}\right)^\alpha\left(\frac{\mathbf{y}_i - c}{2\sigma}\right)^\alpha + \varepsilon(\alpha), \tag{2.115}$$

where the factorial and the length of α are defined, respectively, as $\alpha! = \alpha_1!\alpha_2!\cdots\alpha_d!$ and $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$. The IP can then be written using this form as

$$V_T(\mathbf{y}) \approx \frac{1}{N^2(4\pi\sigma^2)^{d/2}}\sum_{j=1}^{N}\sum_{B}\sum_{\alpha\ge 0} C_\alpha(B)\exp\left(-\frac{\|\mathbf{y}_j - c_B\|^2}{4\sigma^2}\right)\left(\frac{\mathbf{y}_j - c_B}{2\sigma}\right)^\alpha, \tag{2.116}$$

where the coefficients C_α are given by

$$C_\alpha(B) = \frac{2^{|\alpha|}}{\alpha!}\left\{\sum_{e_i\in B}\exp\left(-\frac{\|\mathbf{y}_i - c_B\|^2}{4\sigma^2}\right)\left(\frac{\mathbf{y}_i - c_B}{2\sigma}\right)^\alpha\right\}. \tag{2.117}$$

The coefficients C_α are lexicographically ordered before storage because the expansion of multivariate polynomials can be performed efficiently in this form. For a *d*-dimensional polynomial of order p, all terms are stored in a vector of length

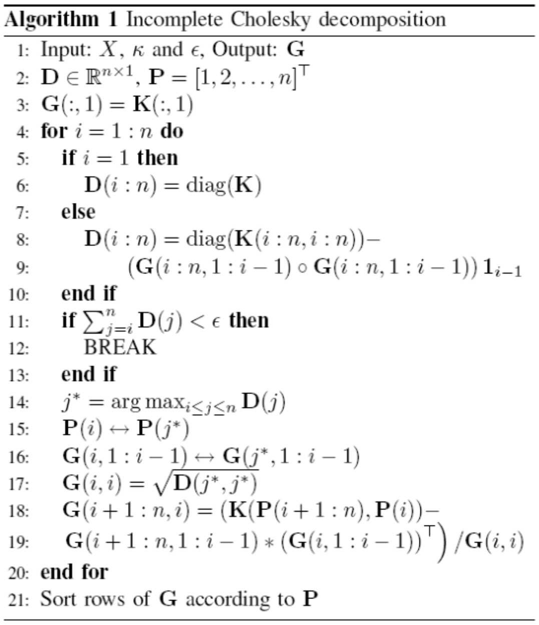
$$r_{p,d} = \binom{p+d}{d} = \frac{(p+d)!}{d!\,p!}.$$

If the series is truncated at order p, then the number of terms is $r_{p,d}$ which is much less than $\Pi_p^d = p^d$ in higher dimensions. The total computational complexity is $O(BNr_{p,d})$, where *B* is the number of clusters.

These algorithms decrease the number of computations appreciably when estimating entropy and divergence in ITL with the Gaussian kernel because the computations become $O(N)$. However, we have to remember that they are not exact evaluations, therefore the number of terms in the expansions and the number of clusters have to be determined appropriately according to the application. Coprocessors for desktop computers in the GigaFlop range have been developed for astrophysics applications [179], but they lack the flexibility required for ITL where the number of dimensions of the problem, the kernel type are free parameters and where the computations are much more general than just evaluating forces.

Incomplete Cholesky Decomposition

Any $N \times N$ symmetric positive definite matrix $\mathbf{K}$ can be expressed as $\mathbf{K} = \mathbf{G}^{\mathrm{T}}\mathbf{G}$ where $\mathbf{G}$ is an $N \times N$ lower triangular matrix with positive diagonal entries. This decomposition is known as the Cholesky decomposition which is a special case of the LU decomposition for a symmetric positive definite matrix [116]. However, if the eigenvalues of $\mathbf{K}$ drop rapidly, then the matrix can be approximated by a $N \times D$ ($D \leq N$) lower triangular matrix $\tilde{\mathbf{G}}$ with arbitrary accuracy; that is, $\left\|\mathbf{K} - \tilde{\mathbf{G}}^T\tilde{\mathbf{G}}\right\| < \varepsilon$ where ε is a small positive number of choice and || · || is a suitable matrix norm. This decomposition is called the incomplete Cholesky decomposition (ICD) [116]. It is observed that in kernel learning [100], depending on the eigenstructure of the matrix, even $D << N$ provides desired accuracy in practice. Although computation involving $\mathbf{K}$ can be largely simplified using $\tilde{\mathbf{G}}$, computing $\tilde{\mathbf{G}}$ itself appears as an overhead, but fortunately there are efficient algorithms to accomplish this task [116]. The particular algorithm in the following table takes a greedy approach and tries to minimize the trace of the residual $\mathbf{K} - \tilde{\mathbf{G}}^T\tilde{\mathbf{G}}$. Its space complexity is $O(ND)$ and the time complexity is $O(ND^2)$, exactly the same complexity as the factorization of $\tilde{\mathbf{G}}$. We provide the algorithm below.



Algorithm 1 Incomplete Cholesky decomposition

```
Input: $X$, $\kappa$ and $\epsilon$, Output: $\mathbf{G}$
$\mathbf{D} \in \mathbb{R}^{n \times 1}$, $\mathbf{P} = [1, 2, \ldots, n]^\top$
$\mathbf{G}(:, 1) = \mathbf{K}(:, 1)$
for $i = 1 : n$ do
    if $i = 1$ then
        $\mathbf{D}(i : n) = \mathrm{diag}(\mathbf{K})$
    else
        $\mathbf{D}(i : n) = \mathrm{diag}(\mathbf{K}(i : n, i : n)) -$
            $(\mathbf{G}(i : n, 1 : i-1) \circ \mathbf{G}(i : n, 1 : i-1))\,\mathbf{1}_{i-1}$
    end if
    if $\sum_{j=i}^{n} \mathbf{D}(j) < \epsilon$ then
        BREAK
    end if
    $j^* = \arg\max_{i \leq j \leq n} \mathbf{D}(j)$
    $\mathbf{P}(i) \leftrightarrow \mathbf{P}(j^*)$
    $\mathbf{G}(i, 1 : i-1) \leftrightarrow \mathbf{G}(j^*, 1 : i-1)$
    $\mathbf{G}(i, i) = \sqrt{\mathbf{D}(j^*, j^*)}$
    $\mathbf{G}(i+1 : n, i) = (\mathbf{K}(\mathbf{P}(i+1 : n), \mathbf{P}(i)) -$
     $\mathbf{G}(i+1 : n, 1 : i-1) * (\mathbf{G}(i, 1 : i-1))^\top\big) / \mathbf{G}(i, i)$
end for
Sort rows of $\mathbf{G}$ according to $\mathbf{P}$
```

Fast Computation of IP

The information potential can be written in terms of a symmetric positive Gram matrix as

$$\hat{V}(X)=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa(x_i-x_j)=\frac{1}{N^2}\mathbf{1}_N^T\mathbf{K}_{XX}\mathbf{1}_N=\frac{1}{N^2}\left\|\mathbf{1}_N^T\tilde{\mathbf{G}}_{XX}\right\|_2^2, \qquad (2.118)$$

where $\mathbf{1}_N$ is a vector of all 1 and size N. The computation decreases from $O(N^2)$ to $O(ND^2)$, and we have obtained precisions of 10^{-6} for 1000 sample datasets, while reducing the computation time 100-fold [292]. The quadratic mutual information algorithms of Eq. (2.97) use only the IP so they can be easily written as

$$\begin{aligned}\hat{I}_{ED}=&\frac{1}{N^2}\mathbf{1}_{\mathbf{D_x}}^{\mathbf{T}}(\tilde{\mathbf{G}}^T{}_{XX}\tilde{\mathbf{G}}_{YY}\circ\tilde{\mathbf{G}}^T{}_{XX}\tilde{\mathbf{G}}_{YY})\mathbf{1}_{\mathbf{D_y}}+\frac{1}{N^4}\left\|\mathbf{1}_N^T\tilde{\mathbf{G}}_{XX}\right\|_2^2\left\|\mathbf{1}_N^T\tilde{\mathbf{G}}_{YY}\right\|_2^2\\&-\frac{2}{N^3}(\mathbf{1}_N^T\tilde{\mathbf{G}}_{XX})(\tilde{\mathbf{G}}^T{}_{XX}\tilde{G}_{YY})(\tilde{\mathbf{G}}_{YY}^T\mathbf{1}_N)\end{aligned} \qquad (2.119)$$

$$\hat{I}_{CS}=\log\frac{\mathbf{1}_{\mathbf{D_x}}^{\mathbf{T}}(\tilde{\mathbf{G}}^T{}_{XX}\tilde{\mathbf{G}}_{YY}\circ\tilde{\mathbf{G}}^T{}_{XX}\tilde{\mathbf{G}}_{YY})\mathbf{1}_{\mathbf{D_y}}\left\|\mathbf{1}_N^T\tilde{\mathbf{G}}_{XX}\right\|_2^2\left\|\mathbf{1}_N^T\tilde{\mathbf{G}}_{YY}\right\|_2^2}{\left((\mathbf{1}_N^T\tilde{\mathbf{G}}_{XX})(\tilde{\mathbf{G}}_{XX}^T\tilde{\mathbf{G}}_{YY})(\tilde{\mathbf{G}}_{YY}^T\mathbf{1}_N)\right)^2}. \qquad (2.120)$$

In these expressions the symbol $\circ$ denotes the elementwise matrix multiplication (Hadamard or Schur product). The computational complexity decreases dramatically from $O(N^3)$ to $O(N(D^2_x + D^2_y + D_xD_y))$.

Fast Computation of the CIP

Unfortunately, the cross information potential does not yield a symmetric positive definite Gram matrix, therefore the above algorithm cannot be directly applied. However, one can augment the matrix to make it symmetric: if the Gram matrix for the CIP is denoted $\mathbf{K}_{XY}$, we create a matrix of double size given by

$$\mathbf{K}_{ZZ}=\begin{vmatrix}\mathbf{K}_{XX} & \mathbf{K}_{XY}\\ \mathbf{K}_{XY} & \mathbf{K}_{YY}\end{vmatrix}.$$

This may seem a waste, but it turns out that in many ITL descriptors each one of the parts of this matrix is needed as we show below. The CIP then can be written as

$$\hat{V}(X,Y)=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa(x_i-y_j)=\frac{1}{N^2}\mathbf{e}_1^T\mathbf{K}_{zz}\mathbf{e}_2=\frac{1}{N^2}\left(\mathbf{e}_1^T\tilde{\mathbf{G}}_{ZZ}\right)\left(\tilde{\mathbf{G}}_{ZZ}\mathbf{e}_2\right), \qquad (2.121)$$

where

$$\mathbf{e}_1=\{\underbrace{1,\ldots,1}_{N},\underbrace{0,\ldots,0}_{N}\}^T \text{ and } \mathbf{e_2}=\{\underbrace{0,\ldots,0}_{N},\underbrace{1,\ldots,1}_{N}\}^T.$$

The computational complexity of the CIP is also $O(ND^2)$. The divergences of Eq. (2.86) and (2.90) that use the CIP can be written in matrix form as

$$\hat{D}_{ED}=\frac{1}{N^2}(\mathbf{e}_1^{\mathbf{T}}\tilde{\mathbf{G}}_{ZZ})(\tilde{\mathbf{G}}_{ZZ}^T\mathbf{e}_1)+\frac{1}{N^2}(\mathbf{e}_2^{\mathbf{T}}\tilde{\mathbf{G}}_{ZZ})(\tilde{\mathbf{G}}_{ZZ}^T\mathbf{e_2})-\frac{2}{N^2}(\mathbf{e}_1^{\mathbf{T}}\tilde{\mathbf{G}}_{ZZ})(\tilde{\mathbf{G}}_{ZZ}^T\mathbf{e_2}) \qquad (2.122)$$

$$\hat{D}_{CS} = \log \frac{(\mathbf{e}_1^{\mathrm{T}}\tilde{\mathbf{G}}_{ZZ})(\tilde{\mathbf{G}}_{ZZ}^{T}\mathbf{e}_1)(\mathbf{e}_2^{\mathrm{T}}\tilde{\mathbf{G}}_{ZZ})(\tilde{\mathbf{G}}_{ZZ}^{T}\mathbf{e}_2)}{\left((\mathbf{e}_1^{\mathrm{T}}\tilde{\mathbf{G}}_{ZZ})(\tilde{\mathbf{G}}_{ZZ}^{T}\mathbf{e}_2)\right)^2}. \tag{2.123}$$

The computational complexity is identical to the CIP. The advantage of the ICD with respect to the FGT is the simpler data structures for the computation, in as much as everything is done in vector matrix products.

2.12 Conclusion

This chapter presented the definition of Renyi's family of entropies, their meaning and relationship with Shannon, and their impact in developing nonparametric estimators for entropy. In particular the argument of the log of quadratic Renyi's entropy called here the information potential, can be estimated directly from data with kernels. The IP can be considered on a par with nonparametric estimators of mean and variance, but unlike them it depends upon one free parameter that needs to be estimated from the data structure and controls the bias and variance for finite datasets. This brings flexibility to the designer, but also requires proper selection. The simple Silverman's rule of density estimation is normally sufficient when data are low-dimensional, but more involved techniques such as cross-validation are required for more accurate results. This dependence on the kernel size makes the IP appropriate primarily for relative comparisons within the same dataset, but when the goal is to optimize a cost function this is perfectly suitable.

From the IP estimator we developed a physical interpretation for the PDF estimated with kernels as a potential field, where samples interact with each other under information forces. From the information potential we proposed two dissimilarity measures in probability spaces that can also be estimated directly from data because they are functions of the IP. One important part of the chapter addresses a detailed treatment of the properties of these estimators in adaptation in as much as this is going to be instrumental for the rest of the chapters, including an analysis of the mean and variance of the IP.

The chapter presents all the necessary equations to implement ITL quantities and they are used extensively throughout the book. The estimator of entropy and ITL divergences are $O(N^2)$ and the estimator for QMI is $O(N^3)$, therefore we also presented two approximations to the computation that provides tremendous speedups in most cases, by using the concept of fast Gauss transform and incomplete Cholesky decomposition.