

# Reproducing Kernel Hilbert Space

$E$ : abstract set.

$\mathcal{H}$ : Hilbert space of functions  $E \mapsto \mathbb{C}$ , equipped with  $\langle \cdot, \cdot \rangle_{\mathcal{H}}: \mathcal{H} \times \mathcal{H} \mapsto \mathbb{C}$ . Associated norm  $\| \cdot \|_{\mathcal{H}}$ :  $\| \phi \|_{\mathcal{H}} = \langle \phi, \phi \rangle_{\mathcal{H}}^{1/2}$ ,  $\phi \in \mathcal{H}$

Evaluation function  $e_t$ ,  $t \in E$ : is a mapping  $\mathcal{H} \mapsto \mathbb{C}$ ,  $g \mapsto e_t(g) = g(t)$ .

Denote the conjugate of  $x$  to be  $\bar{x}$ , the transconjugate of a matrix  $M$  to be  $M^*$ .

Denote  $\mathbb{C}^E$  to be the set of functions  $E \mapsto \mathbb{C}$ .

Example: Let  $\mathcal{H}$  be a finite dimensional vector space of functions, with basis  $(f_1, \dots, f_n)$ . The inner produce on  $\mathcal{H}$  is solely defined by  $g_{ij} = \langle f_i, f_j \rangle$ . If

$$v = \sum_{i=1}^n v_i f_i \quad w = \sum_{i=1}^n w_i f_i$$

then

$$\langle v, w \rangle_{\mathcal{H}} = \sum_{i=1}^n \sum_{j=1}^n v_i \bar{w}_j g_{ij}$$

The matrix  $G = (g_{ij})$  is the Gram matrix,  $G = G^*$ , and  $v^* G v > 0$  when  $v \neq 0$ .

A function

$$K : E \times E \rightarrow \mathbb{C} \\ (s, t) \mapsto K(s, t)$$

is a reproducing kernel of the Hilbert space  $\mathcal{H}$  if and only if

$$(1) \forall t \in E, \quad K(\cdot, t) \in \mathcal{H} \\ (2) \forall t \in E, \forall \phi \in \mathcal{H}, \quad \langle \phi, K(\cdot, t) \rangle = \phi(t)$$

As a consequence,  $\langle K(\cdot, s), K(\cdot, t) \rangle = K(t, s)$ . A Hilbert space that possesses a  $K$  is called *a reproducing kernel Hilbert space*.

For a stationary process, the autocovariance is

$$\gamma(h) = \mathbb{E}[(x_t - \mu)(x_{t-h} - \mu)]$$

independent of  $t$ .

The autocorrelation is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

The spectral density of the stochastic process is the Fourier transform of the autocovariance

$$f(w) = \frac{1}{\sqrt{(2\pi)^n}} \int_{-\infty}^{\infty} \gamma(h) e^{-i\omega h} dh$$

Note the spectral density is a population quantity independent of realization.

*Theorem.* Any finite dimensional Hilbert space of functions has a reproducing kernel

$$K(x, y) = \sum_{i=1}^n e_i(x) \bar{e}_i(y),$$

where  $(e_1, \dots, e_n)$  is an orthogonal basis in  $\mathcal{H}$ , i.e.  $\langle e_i, e_j \rangle_{\mathcal{H}} = \delta_{ij}$ .

*Gauss-Markov Theorem* For a linear regression model, if the errors have (1) zero expectation, and (2) uncorrelated and equal variance, then the *best linear unbiased estimator* of coefficients is the ordinary least squares (OLS) estimator.

Simple Kriging is a linear estimator

$$\hat{Z}(x_0) = m + W^T(Z - m) = m + \sum_{i=1}^N w_i(x_0) (Z(x_i) - m),$$

where  $\mathbb{E}[Z(x)] = m$  is the known mean. The estimation error is

$$\epsilon(x_0) = \hat{Z}(x_0) - Z(x_0)$$

It should satisfy two conditions: 1. unbiased, 2. minimum variance. 1 is automatically satisfied. For 2,

$$\begin{aligned} \text{Var}[\epsilon(x_0)] &= \text{Var}[m + W^T(Z - m) - Z(x_0)] \\ &= \text{Var}\left[\underbrace{(1 - W^T)m}_{\text{Var}=0} + W^T Z - Z(x_0)\right] \\ &= (W^T - 1) \begin{pmatrix} C & c_0 \\ c_0 & c_{00} \end{pmatrix} \begin{pmatrix} W \\ -1 \end{pmatrix} \\ &= W^T C W - 2W^T c_0 + c_{00} \end{aligned}$$

Thus  $W^* = C^{-1}c_0$ , and the minimum estimation variance is  $\text{Var}^*[\epsilon(x_0)] = c_{00} - c_0^T C^{-1} c_0$ .

**Example** Let  $E = \mathbb{R}$ ,  $\mathcal{H} = \{\phi \mid \phi \text{ is continuous, } \phi \text{ and } \phi' \in L^2(\mathbb{R})\}$ . Inner product is defined by

$$\langle \phi, \psi \rangle_{\mathcal{H}} = \int_{\mathbb{R}} (\phi\psi + \phi'\psi') dx$$

Then  $\mathcal{H}$  has the reproducing kernel

$$K(x, y) = \frac{1}{2} \exp(-|x - y|)$$

To verify  $K(x, y)$  is indeed a reproducing kernel for  $\mathcal{H}$ , first we have  $K(\cdot, y) \in \mathcal{H}$ . Second, we verify  $\langle \phi, K(\cdot, y) \rangle_{\mathcal{H}} = \phi(y)$ . We have

$$\frac{\partial}{\partial x} K(x, y) = \begin{cases} -K(x, y) & \text{if } x > y \\ K(x, y) & \text{if } x < y \end{cases}$$

and

$$\frac{\partial^2}{\partial x^2} K(x, y) = K(x, y) \quad \text{if } x \neq y$$

Integration by parts gives

$$\begin{aligned} \langle \phi, K(\cdot, y) \rangle_{\mathcal{H}} &= \int_{\mathbb{R}} \phi(x) K(x, y) dx + \phi(x) K(x, y) \Big|_{-\infty}^y + \phi(x) K(x, y) \Big|_y^{\infty} - \int_{-\infty}^y \phi(x) K(x, y) dx - \int_y^{\infty} \phi(x) K(x, y) dx \\ &= \phi(y) \end{aligned}$$

Thus  $K(x, y)$  is the reproducing kernel of  $\mathcal{H}$ . □

**Positive type function** A function  $K: E \times E \rightarrow \mathbb{R}$  is called a *positive type function* if

$$\forall (x_1, \dots, x_n) \in E^n$$

we have matrix  $K$  defined by  $K(x_i, x_j)$  is positive definite.

**Lemma** Let  $\mathcal{H}$  be a Hilbert space with inner product  $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ . Let  $\phi: E \rightarrow \mathcal{H}$  (arbitrary). Then the function  $K$

$$\begin{aligned} E \times E &\rightarrow \mathbb{R} \\ (x, y) &\mapsto K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} \end{aligned}$$

is of positive type.

**Cauchy-Schwarz** Let  $K$  be any positive type function on  $E \times E$ , then

$$|K(x, y)|^2 \leq K(x, x)K(y, y)$$

*Proof* Let  $\alpha = \frac{K(y, x)}{K(x, x)}$ , and  $z = y - \alpha x$ , we have

$$K(z, x) = K(y - \alpha x, x) = 0$$

Thus,

$$K(y, y) = K(z + \alpha x, z + \alpha x) = K(z, z) + \alpha^2 K(x, x) \geq 0 \square$$

**Moore-Aronszajn Theorem** Let  $K$  be any positive type function on  $E \times E$ . There exists *one and only one* Hilbert space  $\mathcal{H}$  of functions on  $E$  with  $K$  as the reproducing kernel.  $\mathcal{H}_0$  spanned by  $\{K(\cdot, x)_{x \in E}\}$  is a dense subspace of  $\mathcal{H}$ . Further, if  $f = \sum_{i=1}^n K(\cdot, x_i)$ , and  $g = \sum_{j=1}^m \beta_j K(\cdot, y_j)$ , we have

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_i \sum_j \alpha_i \beta_j K(y_j, x_i)$$

The Moore-Aronszajn theorem construct equivalency between positive type functions, reproducing kernel, and reproducing kernel Hilbert space. The next theorem gives equivalency between the definition of a positive type function  $K$  and the definition of a mapping  $T: E \mapsto \text{some space } l^2(X)$ .

**Theorem** A function  $K: E \times E \mapsto \mathbb{R}$  is a reproducing kernel or positive type function, iff there exists a mapping  $T: E \mapsto l^2(X)$  such that

$$\forall (x, y) \in E \times E \quad K(x, y) = \langle T(x), T(y) \rangle_{l^2(X)} = \sum_{\alpha \in X} (T(x))_{\alpha} (T(y))_{\alpha}$$

*Example* Consider  $K(x, y) = \min(x, y)$ ,  $\mathbb{R}^+ \times \mathbb{R}^+ \mapsto \mathbb{R}^+$ . Notice

$$K(x, y) = \int_{\mathbb{R}^+} \mathbf{1}_{[0, y]}(t) \mathbf{1}_{[0, x]}(t) dt = \langle T(y), T(x) \rangle_{L^2_{\mathbb{R}^+}}$$

Thus  $K$  is a reproducing kernel.

**Transformation of kernels** If  $K_1$  is a kernel on  $\mathcal{X}_1$ ,  $K_2$  is a kernel on  $\mathcal{X}_2$ ,  $\alpha > 0$ , and  $A: \mathcal{X}_1 \mapsto \mathcal{X}_2$ , then

- $\alpha K_1$  is a kernel on  $\mathcal{X}_1$ .
- If  $\mathcal{X}_1 = \mathcal{X}_2 \equiv \mathcal{X}$ , then  $K_1 + K_2$  is a kernel on  $\mathcal{X}$ .
- $K_2(A(\cdot), A(\cdot))$  is a kernel on  $\mathcal{X}_1$ .

- $K_1 \times K_2$  (multiplication of real numbers) is a kernel on  $\mathcal{X}_1 \otimes \mathcal{X}_2$ .
- If  $\mathcal{X}_1 = \mathcal{X}_2 \equiv \mathcal{X}$ , then  $K_1 \times K_2$  is a kernel on  $\mathcal{X}$ .

A kernel can be expressed as

$$K(x, x') = \sum_{i=1}^N \sqrt{\lambda_i} e_i(x) \sqrt{\lambda_i} e_i(x'),$$

where  $e_i$  are orthonormal in  $L_2(\mu)$  for a  $\sigma$ -finite measure  $\mu$ :

$$\int_{\mathcal{X}} e_i(x) e_j(x) d\mu(x) = \delta_{ij}$$

Define a Hilbert space  $\mathcal{H}$  to be the space of functions mapping  $\mathcal{X} \mapsto \mathbb{R}$

$$f(x) = \sum_{i=1}^N f_i \sqrt{\lambda_i} e_i(x)$$

Define the projection of  $f$  onto  $e_i(x)$

$$P_i f \equiv f_i = \frac{1}{\sqrt{\lambda_i}} \int_{\mathcal{X}} f(x) e_i(x) d\mu(x),$$

i.e.  $f$  is expressed by a set of characteristic coefficients  $Pf \equiv (P_1 f, \dots, P_N f)^T$ .  $Pf \in \mathbb{R}^N$  is called the *feature space*. Define the inner product of the Hilbert space

$$\langle f, g \rangle_{\mathcal{H}} = (Pf)^T (Pg),$$

which converts the inner product in  $\mathcal{H}$  into inner product in  $\mathbb{R}^N$ .

The evaluation function

$$K(\cdot, x) = \sum_{i=1}^N \sqrt{\lambda_i} e_i(x) \sqrt{\lambda_i} e_i(\cdot) \in \mathcal{H}$$

$$PK(\cdot, x) = \left( \sqrt{\lambda_1} e_1(x), \dots, \sqrt{\lambda_N} e_N(x) \right)^T$$

We can verify

$$K(x, x') = \langle K(\cdot, x), K(\cdot, x') \rangle_{\mathcal{H}} = (PK(\cdot, x))^T (PK(\cdot, x'))$$

A subtle point is  $\{K(\cdot, x) | x \in \mathcal{X}\} \subseteq \mathcal{H}$ .

**Cauchy-Schwarz** Suppose  $\{f_i\}_{i=1}^N$  is square summable, then

$$|f(x)| = \left| \sum_{i=1}^N f_i \sqrt{\lambda_i} e_i(x) \right|$$

$$\leq \sqrt{\sum_{i=1}^N f_i^2} \cdot \sqrt{\sum_{i=1}^N \lambda_i e_i^2(x)} = \|f\|_{\mathcal{H}} \sqrt{K(x, x)}$$

**Theorem** Convergence in Hilbert space norm  $\|f - f_n\|_{\mathcal{H}} \rightarrow 0, n \rightarrow \infty$  implies pointwise convergence  $|f(x) - f_n(x)| \rightarrow 0, n \rightarrow \infty$ . (Proven by Cauchy-Schwarz).

Let  $\mathcal{H}$  be a vector space over field  $F$ , then the space  $\mathcal{H}^*$  consisting of all linear functionals  $\phi : \mathcal{H} \mapsto F$  is the *dual space* of  $\mathcal{H}$ .

The reproducing kernel Hilbert space can also be written as

$$\mathcal{H}(\mathcal{X}) = \text{span}\{K(\cdot, x) : \forall x \in \mathcal{X}\}$$

**Theorem** Suppose  $K(x, y) = \Phi(x - y)$ ,  $\mathcal{X} = \mathbb{R}^n$ ,  $\mathcal{H}$  is the RKHS of  $K$ , and

$$\mathcal{H} \subseteq \{f \mid \frac{\hat{f}}{\sqrt{\hat{\Phi}}} \in L_2(\mathbb{R}^n)\}$$

Then

$$\langle f, g \rangle_{\mathcal{H}} = \frac{1}{\sqrt{(2\pi)^n}} \int_{\mathbb{R}^n} \frac{\hat{f}(w) \bar{\hat{g}}(w)}{\hat{\Phi}(w)} dw,$$

where  $\bar{\cdot}$  is the Fourier transformation.  $\hat{\Phi}(w)$  is the Fourier transformation of  $\Phi(x)$ .

Proof:

$$\begin{aligned} f &= \sum_i f_i K(\cdot, x_i), \quad \hat{f} = \sum_i f_i \hat{\Phi} e^{-iwx_i} \\ g &= \sum_j g_j K(\cdot, y_j), \quad \hat{g} = \sum_j g_j \hat{\Phi} e^{-iwy_j} \\ \text{rhs} &= \frac{1}{\sqrt{(2\pi)^n}} \int_{\mathbb{R}^n} \sum_{ij} f_i g_j \hat{\Phi} e^{-iwx_i - iwy_j} \\ &= \sum_{ij} f_i g_j \left( \frac{1}{\sqrt{(2\pi)^n}} \int_{\mathbb{R}^n} \hat{\Phi} e^{-iwx_i - iwy_j} \right) \\ &= \sum_{ij} f_i g_j \Phi(x_i - y_j) \\ &= \sum_{ij} f_i g_j K(x_i, y_j) \\ &= \langle f, g \rangle_{\mathcal{H}} \quad \square \end{aligned}$$

**Theorem**, If  $K$  is a positive type function,  $\{x_i\}_{i=1}^N$  are distinct points. Then there exist functions  $u_{(j)}^* \in \text{span}\{K(\cdot, x_i), i = 1, \dots, N\}$  such that  $u_{(j)}^*(x_i) = \delta_{ij}$ .

Proof:

$$u_{(j)}^* = \sum_{i=1}^N u_{(j)i}^* K(\cdot, x_i)$$

It can be seen  $u_{(j)i}^* = (K^{-1})_{ij}$ , where  $K_{ij} = K(x_i, x_j)$ .  $u^*$  is called the *cardinal functions* on  $\{x_i\}_{i=1}^N$ .  $\square$

Thus, the interpolant can be written as

$$Pf(x) \equiv \sum_{i=1}^N f(x_i) u_{(i)}^*(x),$$

which is the Kriging estimator.

**Definition** First define:

$$\begin{aligned} Q(x; u, \{x_i\}_{i=1}^N) &= \left\| K(\cdot, x) - \sum_j u_j K(\cdot, x_j) \right\|_{\mathcal{H}}^2 \\ &= K(x, x) + \sum_i \sum_j u_i u_j K(x_i, x_j) - 2 \sum_j u_j K(x, x_j), \end{aligned}$$

where  $u \in \mathbb{R}^n$ .

The **power function** is defined as

$$\left| P_{K, \{x_i\}_{i=1}^N}(x) \right|^2 \equiv Q(x; u^*, \{x_i\}_{i=1}^N),$$

where

$$u^* = u^*(x) = \left( u_{(1)}^*(x), \dots, u_{(N)}^*(x) \right)^T = (K_{ij})^{-1} (K(x, x_i))^T$$

Also,

$$\begin{aligned} \left| P_{K, \{x_i\}_{i=1}^N}(x) \right|^2 &= K(x, x) - \sum_i \sum_j u_i^* K(x_i, x_j) u_j^* \\ &= K(x, x) - \sum_i u_i^* K(x, x_i) \end{aligned}$$

And,

$$\left| P_{K, \{x_i\}_{i=1}^N}(x) \right|^2 = \text{Var}^* [\epsilon(x)]$$

**Theorem** If  $f \in \mathcal{H}$ , then

$$|f(x) - Pf(x)| \leq \underbrace{\left| P_{K, \{x_i\}_{i=1}^N}(x) \right|}_{\text{independent of } f \text{ value}} \|f\|_{\mathcal{H}}$$

**Theorem** Given  $x, \{x_i\}_{i=1}^N$ , i.e. view  $Q$  as only depending on  $u$ . Then

$$\min Q(u) = Q(u^*(x))$$

**Definition** *Fill distance*

$$h = h_{\{x_i\}_{i=1}^N, \mathcal{X}} = \sup_{x \in \mathcal{X}} \min_{x_j \in \{x_i\}_{i=1}^N} \|x - x_j\|_2,$$

i.e. the radius of the largest empty ball placed among the dataset.

**Definition** *Attach*

Given a Gaussian process  $\xi(x)$  with covariance function  $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ , the *RKHS attached* to  $\xi$  is the completion of the linear space of all functions:

$$x \in \mathcal{X} \mapsto \sum_i \alpha_i K(x, x_i), \quad \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}, i \in \mathbb{N}$$

with inner product defined as before (using *evaluation* property).

## 1 Proving Twin Model's Convergence

$\mathcal{X} \subseteq \mathbb{R}^d$  is compact.  $\xi(x)$ : Gaussian process with zero mean, known covariance. Existing samples  $\{x_i\}_{i=1}^n$ , sample values  $\xi(x_i)$ . Maximum value  $M_n \equiv \xi(x_1) \vee \dots \vee \xi(x_n)$ .  $z_+ \equiv \max\{z, 0\}$ . The expected improvement algorithm maximizes

$$\rho_n(x) \equiv \mathbb{E} [(\xi(x) - M_n)_+ | \xi(x_1), \dots, \xi(x_n)]$$

**Theorem** A global optimization algorithm converges for *all* continuous functions iff the sequence of evaluation points produced by the algorithm is dense for *all* continuous functions [Torn and Zilinskas 1989, Theorem 1.3].

The objective function is modeled as  $\xi(x, \omega) : \mathcal{X} \times \Omega \mapsto \mathbb{R}$ , where  $\omega$  is the stochastic dimension. A deterministic optimization strategy maps  $\omega$  to a search sequence in  $\mathcal{X}^{\mathbb{N}}$ :

$$\underline{x}(\omega) \equiv (x_1(\omega), x_2(\omega), \dots),$$

with the property  $x_{n+1}$  depends only on  $\xi(x_1, \omega), \dots, \xi(x_n, \omega)$ .

More formally, the search strategy generates a random sequence  $\underline{x}$  in  $\mathcal{X}$ , where  $x_{n+1}$  is  $\mathcal{F}_n$ -measurable.  $\mathcal{F}_n$  is the  $\sigma$ -algebra generated by  $\xi(x_1, \omega), \dots, \xi(x_n, \omega)$ . The conditional expectation of  $\xi(x)$  given  $\mathcal{F}_n$  is  $\hat{\xi}_n(x; \underline{x}_n)$

$$\hat{\xi}_n(x, \omega; \underline{x}_n) = \sum_{i=1}^n \lambda_n^i(x; \underline{x}_n) \xi(x_i, \omega)$$

$$\sigma_n^2(x; \underline{x}_n) = \mathbb{E}_\omega \left[ \left( \xi(x, \omega) - \hat{\xi}_n(x, \omega; \underline{x}_n) \right)^2 \right]$$

Notice  $\sigma_n^2(x; \underline{x}_n)$  is independent of  $\omega$ .

**Definition** *No-Empty-Ball property*

The covariance  $K(\cdot, \cdot)$  of a Gaussian process  $\xi$  has the *NEB* property if, for  $\forall \underline{x}_n \in \mathcal{X}^n, y \in \mathcal{X}$ , the following assertions are equivalent:

- $y$  is an adherent point of  $\underline{x}_n$
- $\sigma_n^2(y; \underline{x}_n) \rightarrow 0$  as  $n \rightarrow \infty$

The optimization strategy generates

$$\begin{aligned} x_1 &= x_{init} \\ x_{n+1} &= \arg \max_{x \in \mathcal{X}} \mathbb{E} [M_n \vee \xi(x) \mid \mathcal{F}_n] \\ &= \arg \max_{x \in \mathcal{X}} \rho_n(x) \\ &= \arg \max_{x \in \mathcal{X}} \gamma \left( \hat{\xi}_n(x) - M_n, \sigma_n^2(x) \right), \end{aligned}$$

with  $\gamma$  being:

- continuous
- $\forall z \leq 0, \gamma(z, 0) = 0$
- $\forall z \in \mathbb{R}, \forall s > 0, \gamma(z, s) > 0$

**Main Theorem** Assume  $K(\cdot, \cdot)$  has the NEB property.  $\mathcal{H}$  is the RKHS associated with  $K$ . Then for  $\forall x_{init} \in \mathcal{X}$  and  $\forall \xi \in \mathcal{H}$ ,  $\underline{x}_n$  generated by the above optimization strategy is dense in  $\mathcal{X}$ .

**Lemma A** Let  $\{x_n\}_{n \geq 1}$  be a sequence in  $\mathcal{X}$  ( $\{x_n\}_{n \geq 1}$  does not need to be generated by EI). Let  $\{y_n\}_{n \geq 1}$  be a convergent sequence in  $\mathcal{X}$  converging to  $y^*$ . Moreover, assume  $\xi$  is a stochastic process satisfying the NEB property. Then the following three conditions are equivalent:

- $y^*$  is an adherent point of  $\{x_n\}_{n \geq 1}$ ,
- $\sigma^2(y_n; \underline{x}_n) \rightarrow 0$  as  $n \rightarrow \infty$ ,
- For  $\forall \xi \in \mathcal{H}$ , we have  $\hat{\xi}_n(y_n, w; \underline{x}_n) \rightarrow \xi(y^*, w)$  as  $n \rightarrow \infty$ .

**Lemma B** Let  $K$  be the covariance of a stationary process in  $\mathbb{R}^n$  and its spectrum be  $\hat{K}(u)$  as  $u \rightarrow \infty$ , assume  $\hat{K}(u) = \Theta(\|u\|^{-2\nu-n})$  with  $0 < \nu < \infty$ ; and let  $\mathcal{H}$  be the RKHS generated by  $K$ . Then <1> for  $\forall x^* \in \mathbb{R}^n$  with  $U \subseteq \mathbb{R}^n$  being a compact neighborhood of  $x^*$ , there exists  $\xi \in \mathcal{H}$  such that  $\text{supp} \xi \subseteq U$  and  $\xi(x^*) > 0$ . <2>  $K$  has the NEB property.

*Proof:* To prove <1> of Lemma B, we use two lemmas:

**Lemma** If  $\nu < \infty$ ,  $\mathcal{H}(\mathbb{R}^n)$  is equivalent to the Sobolev space  $W^{\nu+d/2,2}(\mathbb{R}^n)$ . [Lemma 3, Adam D. Bull, 2011]

**Lemma**  $C_c^\infty(\mathbb{R}^n)$  is dense in  $W^{m,2}(\mathbb{R}^n)$  where  $m > 0$  and  $C_c^\infty(\mathbb{R}^n)$  are  $C^\infty$  functions with compact support. [Lemma 5.1, Ralph E. Showalter, 2010]

(Still can't understand the meaning of *equivalence*). I should be able to get:

$C_c^\infty(\mathbb{R}^n)$  is dense in  $\mathcal{H}(\mathbb{R}^n)$ . Hereby <1> in the lemma.

Then we prove Lemma A.

(i)  $\rightarrow$  (ii) Assume  $y^* \notin \{x_n, n > 1\}$ . Let  $\{x_{\phi_k}\}_k$  be a subsequence of  $\{x_n\}$  converging to  $y^*$ . Let  $\psi_n = \max\{\phi_k; \phi_k \leq n\}$ . Then

$$\sigma_n^2(y_n; \underline{x}_n) = \text{var} \left[ \xi(y_n) - \hat{\xi}_n(y_n; \underline{x}_n) \right] \leq \text{var} \left[ \xi(y_n) - \xi(x_{\psi_n}) \right]$$

using the fact that the Kriging estimator is the best linear unbiased estimator.

As  $x_{\psi_n} \rightarrow y^*$ , and  $K$  is continuous, we have

$$\text{var} [\xi(y_n) - \xi(x_{\psi_n})] = K(y_n, y_n) + K(x_{\psi_n}, x_{\psi_n}) - 2K(y_n, x_{\psi_n}) \rightarrow 0$$

Notice  $\sigma_n^2(x; \underline{x}_n) \equiv \left| P_{K, \{x_i\}_{i=1}^N}(x) \right|^2$

(ii)  $\rightarrow$  (iii) Using Cauchy-Schwarz inequality

$$\left| \xi(y_n) - \hat{\xi}_n(y_n; \underline{x}_n) \right| \leq \left| P_{K, \{x_i\}_{i=1}^N}(y_n) \right| \cdot \|\xi\|_{\mathcal{H}}$$

and continuity of  $\xi$ , we have triangular inequality

$$\left| \hat{\xi}_n(y_n; \underline{x}_n) - \xi(y^*) \right| \leq \left| \hat{\xi}_n(y_n; \underline{x}_n) - \xi(y_n) \right| + |\xi(y_n) - \xi(y^*)| \rightarrow 0$$

as  $n \rightarrow \infty$  for  $\forall \xi \in \mathcal{H}$ .

(iii)  $\rightarrow$  (i) Suppose this conclusion does not hold, then there exists a bounded neighborhood  $U$  of  $y^*$  which does not intersect  $\{x_i\}_{i=1}^\infty$ . Using  $\langle 1 \rangle$  of Lemma B, we can construct  $\xi \in \mathcal{H}$  compactly supported in  $U$ , and  $\xi(y) = 1$ . Thus  $\hat{\xi}_n(y; \underline{x}_n) = 0$ . This violates (iii). Thus completes the proof of Lemma A.

Lemma A establishes the equivalence of (i)  $\leftrightarrow$  (iii). Thus  $K$  satisfies the NEB property. This completes the proof of Lemma B.  $\square$

**Lemma C** For  $\forall \xi \in \mathcal{H}$ ,

$$\lim_{n \rightarrow \infty} \inf_n \gamma(\hat{\xi}_n(x_{n+1}) - M_n, \sigma_n^2(x_n)) = 0$$

*Proof:* Assume  $y^*$  is a cluster point of  $\{x_n\}$ , and  $\{x_{\phi_n}\}$  be a subsequence of  $\{x_n\}$

*AIAA 2002-0317 Using gradient to construct cokriging approximation Hyong-Seog Chung*

Assume  $f(x) \in \mathcal{H}_K \in C^1$ . In addition to sampling  $f(x)$ , assume we also sample  $\nabla f(x)$ . But the sample of gradient is noisy, i.e.  $\widehat{\nabla f(x)} = \nabla f(x) + \eta$ , where  $\eta \sim \mathcal{N}(0, \epsilon^2)$ . Let  $\{x_D\}$  be the sampled points. Therefore

$$\begin{pmatrix} f(x) \\ f(x_D) \\ \widehat{\nabla f(x_D)} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} 0 \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} K(x, x) & K(x, x_D) & K(x, \nabla x_D) \\ K(x, x_D)^T & K(x_D, x_D) & K(x_D, \nabla x_D) \\ K(x, \nabla x_D)^T & K(x_D, \nabla x_D)^T & K(\nabla x_D, \nabla x_D) + \epsilon^2 \end{pmatrix} \right)$$

Define

$$\begin{aligned} s &= \begin{pmatrix} K(x, \nabla x_D) \\ K(x_D, \nabla x_D) \end{pmatrix} \\ L &= \begin{pmatrix} K(x, x) & K(x, x_D) \\ K(x, x_D) & K(x_D, x_D) \end{pmatrix} \\ P &= K(\nabla x_D, \nabla x_D) \end{aligned}$$

Conditioned on  $\eta$ , we have

$$\begin{pmatrix} f(x) \\ f(x_D) \end{pmatrix} \Big| \widehat{\nabla f(x_D)} \sim \mathcal{N} \left( s(P + \epsilon^2)^{-1} \widehat{\nabla f(x_D)}, L - s(P + \epsilon^2)^{-1} s^T \right)$$

Suppose  $\max_{x_i \in \{x_D\}} \|x_i - x\| < \delta$ , then  $\|s\|_{L_2} < \sqrt{n}C\delta$  where  $C$  depends only on  $K$  (need to polish). Also,  $(P + \epsilon^2)^{-1}$  is bounded because  $P$  is a positive definite and  $\epsilon^2 > 0$  is fixed. Therefore, as  $\max_{x_i \in \{x_D\}} \|x_i - x\| \rightarrow 0$ , we have

$$\begin{pmatrix} f(x) \\ f(x_D) \end{pmatrix} \Big| \widehat{\nabla f(x_D)} \sim \begin{pmatrix} f(x) \\ f(x_D) \end{pmatrix}$$



In other words, the role from gradient sampling is negligible when the sampling is dense. Then we can show one direction of the NEB property.

Using the NEB property of  $\mathcal{H}$ , we can choose  $x^*$  and  $\{x_i\}$  where  $|x_i - x^*| > \delta > 0$ . A function  $f \in \mathcal{H}$  can be constructed to satisfy  $f(x_i) = 0$ . Further, we can choose  $\eta = 0$ . Thus  $\hat{f}(x^*) = 0$ . This should be useful proving (iii) to (i).

For simplicity we assume  $\mathcal{X} = \mathbb{R}$ . Suppose the samplings are  $\{f(y_i)\}_{i=1}^{2N}$ , where  $\{y_i\}_{i=1}^{2N} = \{\{x_i\}_{i=1}^N, \{x_i + \delta\}_{i=1}^N\}$ . Assume the samplings have no noise. The covariance matrix of the samplings is

$$K = \begin{pmatrix} K(\{x\}_{i=1}^N, \{x\}_{i=1}^N) & K(\{x\}_{i=1}^N, \{x\}_{i=1}^N + \delta) \\ K(\{x\}_{i=1}^N + \delta, \{x\}_{i=1}^N) & K(\{x\}_{i=1}^N + \delta, \{x\}_{i=1}^N + \delta) \end{pmatrix}$$

We can construct cardinal functions on  $\{y_i\}_{i=1}^{2N}$ :

$$u_{(i)}^* = \text{span}\{K(\cdot, y_j), j = 1, \dots, 2N\}, i = 1, \dots, 2N$$

such that  $u_i^*(y_j) = \delta_{ij}$ , i.e.

$$u_{(j)}^*(\cdot) = \sum_{i=1}^{2N} u_{(j)i}^* K(\cdot, y_i),$$

with

$$u_{(j)i}^* = (K^{-1})_{ij} = (K^{-1})_{ji}$$

Define

$$Q = \begin{pmatrix} I_N & \\ -\frac{I_N}{\delta} & \frac{I_N}{\delta} \end{pmatrix}$$

and

$$M = \begin{pmatrix} K(\{x\}_{i=1}^N, \{x\}_{i=1}^N) & \nabla_2 K(\{x\}_{i=1}^N, \{x\}_{i=1}^N) \\ \nabla_1 K(\{x\}_{i=1}^N, \{x\}_{i=1}^N) & \nabla_1 \nabla_2 K(\{x\}_{i=1}^N, \{x\}_{i=1}^N) \end{pmatrix},$$

where  $\nabla_k$  means taking the derivative with respect to the  $k$ th entry. We have

$$M = QKQ^T$$

when  $\delta \rightarrow 0$ . For  $f \in \mathcal{H}_K$ , we have the interpolant of  $f$  on the dataset  $\{f(y_i)\}_{i=1}^{2N}$  to be

$$Pf = \sum_{i=1}^{2N} f(y_i) u_{(i)}^*(\cdot) = \sum_{i=1}^{2N} \sum_{j=1}^{2N} f(y_i) u_{(i)j}^* K(\cdot, y_j)$$

Clearly  $Pf \in \mathcal{H}_K$ . Therefore,

$$\begin{aligned} |f(x) - Pf(x)| &= \left\langle f, K(\cdot, x) - \sum_{i=1}^{2N} K(\cdot, y_i) u_{(i)}^*(x) \right\rangle_{\mathcal{H}} \\ &\leq \|f\|_{\mathcal{H}} \left\| K(\cdot, x) - \sum_{i=1}^{2N} K(\cdot, y_i) u_{(i)}^*(x) \right\|_{\mathcal{H}} \\ &= \left\| K(\cdot, x) - \underbrace{\sum_{i=1}^{2N} \sum_{j=1}^{2N} (K^{-1})_{ji} K(x, y_j) K(\cdot, y_i)}_{s_i(x)} \right\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &= \left[ K(x, x) - 2 \sum_{i=1}^{2N} s_i(x) K(x, y_i) + \sum_{i=1}^{2N} \sum_{j=1}^{2N} s_i(x) K(y_i, y_j) s_j(x) \right] \|f\|_{\mathcal{H}} \end{aligned}$$

Define

$$d_j(x) = \sum_{i=1}^{2N} Q_{ji} K(x, y_i) = \begin{pmatrix} K(x, \mathbf{x}) \\ \nabla_2 K(x, \mathbf{x}) \end{pmatrix},$$

where  $\mathbf{x}$  denotes the vector  $(x_1, \dots, x_N)^T$ . Let  $\mathbf{d}(x)$  be a vector whose entries are  $d_i(x)$ , and  $\mathbf{s}(x)$  be a vector whose entries are  $s_i(x)$ . We have

$$s(x) = Q^T M^{-1} d(x)$$

Therefore,

$$\begin{aligned} |f(x) - Pf(x)| &\leq (K(x, x) - 2\mathbf{d}^T(x)M^{-1}\mathbf{d}(x) + \mathbf{d}^T(x)M^{-1}\mathbf{d}(x)) \|f\|_{\mathcal{H}} \\ &= (K(x, x) - \mathbf{d}^T(x)M^{-1}\mathbf{d}(x)) \|f\|_{\mathcal{H}} \\ &= \sigma_n^2(x) \|f\|_{\mathcal{H}}, \end{aligned}$$

where  $\sigma_n^2(x)$  is the posterior variance conditioned on exact samplings of  $f(x_i)$  and  $\nabla f(x_i)$ ,  $i = 1, \dots, N$ . We also have

$$Pf = \sum_{i=1}^N \beta_i^1 f(x_i) + \sum_{i=1}^N \beta_i^2 \nabla f(x_i),$$

and

$$\begin{pmatrix} \beta^1 \\ \beta^2 \end{pmatrix} = \begin{pmatrix} K(\mathbf{x}, \mathbf{x}) & \nabla_2 K(\mathbf{x}, \mathbf{x}) \\ \nabla_1 K(\mathbf{x}, \mathbf{x}) & \nabla_1 \nabla_2 K(\mathbf{x}, \mathbf{x}) \end{pmatrix}^{-1} \begin{pmatrix} K(x, \mathbf{x}) \\ \nabla_2 K(x, \mathbf{x}) \end{pmatrix}$$

Suppose the collocated  $\nabla f(x)$  are sampled with noise  $\eta(x)$ .  $\eta(x)$  is a stochastic process and

$$\text{cov}[f(x), \eta(x)] = 0$$

We model  $\eta(x)$  as a realization of the centered stochastic process with covariance  $H(\cdot, \cdot)$ . The best linear unbiased estimator is given by

$$\hat{P}f = \sum_{i=1}^N \hat{\beta}_i^1 f(x_i) + \sum_{i=1}^N \hat{\beta}_i^2 \widehat{\nabla f(x_i)},$$

where  $\widehat{\nabla f(x_i)}$  indicates noisy gradient sample, and

$$\begin{pmatrix} \hat{\beta}^1 \\ \hat{\beta}^2 \end{pmatrix} = \begin{pmatrix} K(\mathbf{x}, \mathbf{x}) & \nabla_2 K(\mathbf{x}, \mathbf{x}) \\ \nabla_1 K(\mathbf{x}, \mathbf{x}) & \nabla_1 \nabla_2 K(\mathbf{x}, \mathbf{x}) + H(\mathbf{x}, \mathbf{x}) \end{pmatrix}^{-1} \begin{pmatrix} K(x, \mathbf{x}) \\ \nabla_2 K(x, \mathbf{x}) \end{pmatrix}$$

Tried:

- $|f - \hat{P}f|$  triangular inequality
- Prove if  $\sigma^2 \rightarrow 0$ , it have to be  $\sigma_{\text{using just exact samples}}^2 \rightarrow 0$ .

*Thoughts* Suppose we sample a noisy  $f$ :  $\hat{f} = f + \eta$ .  $f \sim \mathcal{H}_K$ ,  $\eta \sim \mathcal{H}_H$ . Clearly no estimator can approach  $f(x)$  using datasets of  $\hat{f}$ , no matter how dense we sample. The best linear unbiased estimator is

$$f_{\text{est}}(x) = k^T (K + H)^{-1} \hat{f}$$

$\hat{f}$  is the noisy dataset.  $K, H$  are covariance matrices of the dataset. The estimation error variance is

$$\sigma^2(x) = K(x, x) - k^T (K + H)^{-1} k$$

Can we show  $\sigma^2(x)$  can never go to 0? Let's define

$$S = K(K^{-1} + H^{-1})K = K + KH^{-1}K,$$

then by Woodbury matrix identity, we have

$$\sigma^2(x) = \underbrace{K(x, x) - k^T K^{-1} k + k^T S^{-1} k}_{\geq 0}$$

We need to find a lower bound of  $k^T S^{-1} k$ .

## 1.1 Cauchy Inequality

Now we prove the Cauchy inequality for sampling with noisy gradient and exact function value. First, the functions  $u$  and  $1 - u$  for  $0 \leq u \leq 1$  belongs to a reproducing kernel hilbert space  $\mathcal{H}_u$ . For example, we can choose a kernel  $G : [0, 1] \times [0, 1] \mapsto \mathbb{R}$ ,  $G(u, v) = |u - v|$ . Assume the function  $f \in \mathcal{H}$  with kernel  $K$ . Then the gradient of  $f \in \mathcal{H}'$ , and is independent of  $f$ . We have  $\mathcal{H} \in \mathcal{H}'$  (Kondrachov embedding theorem). Assume the sample noise  $\eta$ ,  $\text{cov}(\eta(x), \eta(y)) = H(x, y)$ . Construct function

$$F(x, u) = (1 - u)f(x) + u \left[ \frac{\partial f}{\partial x}(x) + \eta(x) \right]$$

For  $x \in \mathbb{R}^n$ ,  $n > 1$ , the definition is

$$F(x, u) = (1 - u)f(x) + u \mathbf{1}^T \left[ \frac{\partial f}{\partial x}(x) + \eta(x) \right]$$

For simplicity we just consider  $n = 1$ .

We have  $F(x, u) \in \mathcal{H}_F$ , with kernel  $K_F((\cdot, \cdot), (x, u)) = K'(\cdot, x)G(\cdot, u)$ . The sampled function values are  $f(x) = F(x, 0)$ , the sampled noisy gradient is  $\frac{\partial f}{\partial x}(x) + \eta(x)$ . For notation simplicity we write the tuple  $(x, u)$  interchangeably with  $xu$ . Denote the sampled data  $\mathbf{y} = \{F(\mathbf{x}, 0), F(\mathbf{x}, 1)\}$ . Apply Cauchy-Schwarz inequality to  $|F(x, 0) - PF(x, 0)|$ , we get

$$|F(x, 0) - PF(x, 0)| \leq \left[ K_F(x0, x0) - \sum_i \sum_j s_i(x0) K_F(\mathbf{y}, \mathbf{y}) s_j(x0) \right] \|F\|_{\mathcal{H}_F}$$

Notice

$$\begin{aligned} K_F(x0, x0) &= K(f(x), f(x)) \\ K_F(\mathbf{y}, \mathbf{y}) &= \begin{pmatrix} K(\mathbf{x}, \mathbf{x}) & \nabla_2 K(\mathbf{x}, \mathbf{x}) \\ \nabla_1 K(\mathbf{x}, \mathbf{x}) & \nabla_1 \nabla_2 K(\mathbf{x}, \mathbf{x}) + H(\mathbf{x}, \mathbf{x}) \end{pmatrix} \\ s_i(x0) &= \sum_j (K_F^{-1})_{ji} K(x0, \mathbf{y}) \end{aligned}$$

Also, because  $0 \leq u \leq 1$ , we have

$$\|F\|_{\mathcal{H}_F} \leq \|f\|_{\mathcal{H}} + \left\| \frac{\partial f}{\partial x} \right\|_{\mathcal{H}'} + \|\eta\|_{\mathcal{H}'}$$

Therefore,

$$\|f(x) - Pf(x)\| \leq \left( \|f\|_{\mathcal{H}} + \left\| \frac{\partial f}{\partial x} \right\|_{\mathcal{H}'} + \|\eta\|_{\mathcal{H}'} \right) \sigma^2$$

where  $\sigma^2$  is the posterior variance of  $f(x)$ .

Use theorem 4.1 from RKHS book chp1.pdf to complete the proof.

**Power set**  $X^Y$  is the set of all functions from  $Y$  to  $X$ .  $2^\Omega$  is the set of all functions from  $\Omega$  to  $\{0, 1\}$ .

**Set function** is a function whose input is a set. The output is usually a number.

**Submodular set function** is a set function  $f : 2^\Omega \rightarrow \mathbb{R}$ , which satisfies the following condition: For every  $X, Y \subseteq \Omega$  with  $X \subseteq Y$ , and every  $x \in \Omega \setminus Y$ , we have  $f(X \cup \{x\}) - f(x) \geq f(Y \cup \{x\}) - f(Y)$ .

## 2 Multi-armed bandit and UCB optimization

Maximization

$$x^* = \arg \max_{x \in D} f(x)$$

Regret

$$r_t = f(x^*) - f(x_t)$$

Cumulative regret after  $T$  rounds

$$R_T = \sum_{t=1}^T r_t$$

We consider function  $f$  with domain  $\mathcal{D}$ . The samples can be taken on a finite point set  $D$ . Let  $f_D$  be  $f$  restricted to  $D$ .

Observation type 1:

$$y_t = f(x_t) + \epsilon_t,$$

Observation type 2:

$$z_t = \nabla f(x_t) + g(x_t),$$

where  $t$  is the round of sample.  $f \in \mathcal{H}_{K_f}$ ,  $g \in \mathcal{H}_{K_g}$ .  $\epsilon_t$  are i.i.d Gaussian noise with zero mean and  $\sigma^2$  variance. For simplicity, assume  $\mathcal{D} \subset \mathbb{R}^1$ .

Let the sampled points be  $\mathbf{x}_T$ . The mutual information between  $f$  and the observations  $\mathbf{y}_T, \mathbf{z}_T$  is:

$$I(\{\mathbf{y}_T, \mathbf{z}_T\}; f) = H(\{\mathbf{y}_T, \mathbf{z}_T\}) - H(\{\mathbf{y}_T, \mathbf{z}_T\} | f)$$

According to

$$I(X, Y) = H(X) + H(Y) - H(X, Y),$$

and

$$I(\{\mathbf{y}_T, \mathbf{z}_T\} | f) = 0$$

we have

$$I(\{\mathbf{y}_T, \mathbf{z}_T\}; f) = H(\{\mathbf{y}_T, \mathbf{z}_T\}) - H(\mathbf{y}_T | f) - H(\mathbf{z}_T | f)$$

Define

$$\Sigma = \begin{pmatrix} K_f + \sigma^2 I_T & (\nabla K_f)^T \\ \nabla K_f & \nabla^2 K_f + K_g \end{pmatrix},$$

where  $K_f, K_g$  are the matrix corresponding to  $\mathbf{x}_T$ . Then

$$I(\{\mathbf{y}_T, \mathbf{z}_T\}; f) = \frac{1}{2} \ln |2\pi e \Sigma| - \frac{1}{2} \ln |2\pi e K_g| - \frac{T}{2} \ln(2\pi e \sigma^2)$$

The mutual information is independent of sampled values.

Next, we try to find the relationship between  $\sigma_{t-1}(x_t)$  and  $\tilde{\sigma}_{t-1}(x_t)$ , where  $\sigma$  indicates the predictive variance of  $f(x_t)$  by using  $\mathbf{y}_{t-1}$  alone, and  $\tilde{\sigma}$  indicates the predictive variance by using  $\mathbf{y}_{t-1}$  and  $\mathbf{z}_{t-1}$ . Let  $k_f \equiv K_f(x_t, \mathbf{x}_{t-1})$ , and  $\nabla k_f \equiv K_f(x_t, \nabla \mathbf{x}_{t-1})$ . We have

$$\tilde{\sigma}_{t-1}^2(x_t) = K_f(x_t, x_t) - (k_f \quad \nabla k_f) \begin{pmatrix} K_f + \sigma^2 I_{T-1} & (\nabla K_f)^T \\ \nabla K_f & \nabla^2 K_f + K_g \end{pmatrix}^{-1} \begin{pmatrix} k_f & \nabla k_f \end{pmatrix}^T$$

$$\sigma_{t-1}^2(x_t) = K_f(x_t, x_t) - k_f (K_f + \sigma^2 I_{T-1})^{-1} k_f^T$$

Define

$$\begin{aligned} A &= K_f + \sigma^2 I_{T-1} \\ B &= \nabla K_f \\ C &= \nabla^2 K_f + K_g \end{aligned}$$

Using the formula

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}^{-1} = \begin{pmatrix} I & -A^{-1}B \\ & I \end{pmatrix} \begin{pmatrix} A^{-1} & \\ & (C - B^T A^{-1} B)^{-1} \end{pmatrix} \begin{pmatrix} I & \\ -B^T A^{-1} & I \end{pmatrix},$$

we get

$$\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} = x^T A^{-1} x + (y - B^T A^{-1} x)^T (C - B^T A^{-1} B)^{-1} (y - B^T A^{-1} x) \geq x^T A^{-1} x$$

Notice  $\Sigma^{-1}$  and  $K_f + \sigma^2 I_{T-1}$  is positive-definite, we have

$$\tilde{\sigma}_{t-1}^2(x_t) \leq \sigma_{t-1}^2(x_t)$$

Using the same reasoning of Lemma 5.1 and 5.2 of Niranjan's paper, we get

$$\tilde{r}_t \leq 2\beta_t^{1/2} \tilde{\sigma}_{t-1}(x_t) \leq 2\beta_t^{1/2} \sigma_{t-1}(x_t),$$

where the sequence  $\{x_1, \dots, x_t\}$  are generated using the estimated gradient. Lemma 5.3 and 5.4 of Niranjan's paper states

$$I(\mathbf{y}_T; \mathbf{f}_T) = \frac{1}{2} \sum_{t=1}^T \log(1 + \sigma^{-2} \sigma_{t-1}^2(x_t))$$

and

$$I(\mathbf{y}_T; \mathbf{f}_T) \leq \gamma_T = \max_{A \subset D: |A|=T} I(\mathbf{y}_A; \mathbf{f}_A)$$

Therefore, for finite  $D$ , twin-model-based GP-UCB optimization is not slower than black-box GP-UCB optimization.

### 3 Information theory

The entropy rate of a stochastic process is the average density of information in a stochastic process. For stochastic process with a countable set, the entropy rate is

$$H(X) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, X_2, \dots, X_n)$$

A stochastic process is mean-ergodic if

$$\frac{1}{T} \int_0^T X(t) dt \rightarrow \mu \quad \text{as } T \rightarrow \infty,$$

where  $\mu$  is the ensemble average. Similarly, a stochastic process is autocovariance-ergodic if

$$\frac{1}{T} \int_0^T [X(t+\tau) - \mu][X(t) - \mu] dt \rightarrow r(\tau) \quad \text{as } T \rightarrow \infty$$

Asymptotic equipartition property (AEP):

$$\mathbb{P} \left[ \lim_{n \rightarrow \infty} -\frac{1}{n} \log p(X_1, \dots, X_n) = H(X) \right] = 1$$

The differential entropy is the extension of Shannon entropy to random variables with continuous distributions (with probability density  $f(X)$ )

$$h(X) = - \int_{\mathbb{X}} f(X) \log f(X) dx$$

The entropy for so-called measure-preserving dynamic systems: Kolmogorov-Sinai-entropy. KSE measures the unpredictability of a dynamical system.

Intuition of Shannon entropy: the number of questions that need to ask to restore the message.

Conditional entropy: the average of  $H(Y|X = x)$  over all possible values  $x$  may take, it measures the expected entropy of  $Y$  if  $X$  has been received beforehand.

$$H(Y|X) = - \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)}$$

Spectral distribution function  $F(\lambda)$  of a covariance function  $\gamma_n$  (for discrete stochastic process)

$$\gamma_n = \int_{-\pi}^{\pi} e^{in\lambda} dF(\lambda),$$

where  $F(-\pi) = 0, F(\pi) = \gamma_0$ .

$$F(\lambda) = \int_{-\pi}^{\lambda} f(t) dt,$$

$f(t)$  is called the spectral density function (SDF).