Energy Distance, Scoring Rules and f-Divergence

Junyi Liao

1 Energy Distance

1.1 Semimetric and Conditionally Negative Definite Functions

Definition 1.1 (Semimetric). Let $\mathcal{X}$ be a nonempty set. Then a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a semimetric on $\mathcal{X}$ if it satisfies the following conditions:

(i) (Nonnegativity). $d(x, x') \geq 0 \ \forall x, x' \in \mathcal{X}$, and $d(x, x') = 0$ if and only if $x = x'$;

(ii) (Symmetry). $d(x, x') = d(x', x) \ \forall x, x' \in \mathcal{X}$.

Moreover, $(\mathcal{X}, d)$ is called a semimetric space.

Note that here we do not assume the triangle inequality for semimetric d. If d satisfies the triangle inequality, i.e., $\forall x, y, z \in \mathcal{X}$, we have $d(x, y) + d(y, z) \geq d(x, z)$, then d is a metric on $\mathcal{X}$, and $(\mathcal{X}, d)$ is called a metric space.

Definition 1.2 (Conditionally negative definite). In a semimetric space $(\mathcal{X}, d)$, d is said to be conditionally negative definite, if $\forall n \in \mathbb{N},\ x_1, \cdots, x_n \in \mathcal{X}$ and $\alpha_1, \cdots, \alpha_n \in \mathbb{R}$ with $\sum_{j=1}^{n} \alpha_j = 0$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j d(x_i, x_j) \leq 0. \tag{1.1}$$

Moreover, d is said to be strictly conditionally negative definite if the inequality (1.1) is strict whenever $x_1, \cdots, x_n$ are distinct and at least one of $\alpha_1, \cdots, \alpha_n$ does not vanish.

Proposition 1.3. Let $(\mathcal{X}, d)$ be a semimetric space. Then d is conditionally negative definite if and only if there exists a Hilbert space $\mathcal{H}$ and an injective map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$d(x, x') = \|\phi(x) - \phi(x')\|_{\mathcal{H}}^2. \tag{1.2}$$

By proposition 3, we immediately know that if d is a conditionally negative definite semimetric, then its square root $d^{1/2}$ must be a metric.

Proposition 1.4. Let $(\mathcal{X}, d)$ be a semimetric space. If d is conditionally negative definite, i.e., d satisfies (1.1), then so does d^q for $0 < q < 1$.

1.2 Energy Distance

Definition 1.5 (Energy distance). Suppose P and Q are two probability measures on $\mathbb{R}^d$ with finite first moments. Then the energy distance between P and Q is defined as

$$D_e(P, Q) := 2\mathbb{E}\|X - Y\|_2 - \mathbb{E}\|X - X'\|_2 - \mathbb{E}\|Y - Y'\|_2, \tag{1.3}$$

where $X, X' \overset{\text{i.i.d.}}{\sim} P$ and $Y, Y' \overset{\text{i.i.d.}}{\sim} Q$.

We denote the characteristic function of X and Y as $\widehat{\mu}_P$ and $\widehat{\mu}_Q$, then their energy distance admits the following representation:

Proposition 1.6. Suppose P and Q are two probability measures on $\mathbb{R}^d$ with finite first moments. The following statements are true:

(i) Let $\widehat{\mu}_P(t) = \int \mathrm{e}^{i\langle t,x\rangle}\mathrm{d}P(x)$, $\widehat{\mu}_Q(t) = \int \mathrm{e}^{i\langle t,x\rangle}\mathrm{d}Q(x)$, then

$$D_e(P,Q) = \frac{1}{C_d}\int_{\mathbb{R}^d} \frac{\left|\widehat{\mu}_P(t) - \widehat{\mu}_Q(t)\right|^2}{\|t\|_2^{d+1}}\mathrm{d}t, \text{ where } C_d = \frac{\pi^{\frac{d+1}{2}}}{\Gamma\left(\frac{d+1}{2}\right)}. \tag{1.4}$$

(ii) $D_e(P,Q) \geq 0$, and $D_e(P,Q) = 0$ if and only if $P = Q$.

The following proposition establishes the equivalence between the energy distance and the L_2-discrepancy (Cramér distance) in univariate case.

Proposition 1.7. Suppose P and Q are two probability measures on $\mathbb{R}$ with finite first moments. Then

$$D_e(P,Q) = 2\int_{-\infty}^{+\infty} \left(P(x) - Q(x)\right)^2 \mathrm{d}x. \tag{1.5}$$

Inspired by equation (1.3), we can replace the distance function $\|\cdot - \cdot\|$ with other metric or semimetric d. To ensure the existence of such distances, we need to define the class of measures having finite θ-moments with respect to some semimetric d:

$$\mathcal{M}_d^\theta(\mathcal{X}) = \left\{\nu \text{ is a finite signed measure on } \mathcal{X} : \ \exists x_0 \in \mathcal{X} \text{ such that } \int d^\theta(x,x_0)\mathrm{d}|\nu|(x) < \infty\right\} \tag{1.6}$$

Definition 1.8 (Generalized energy distance). Let $(\mathcal{X}, d)$ be a semimetric space. Suppose $P, Q \in \mathcal{M}_d^1(\mathcal{X})$. Then the energy distance between P and Q with respect to d is defined as

$$D_{e,d}(P,Q) := 2\mathbb{E}[d(X,Y)] - \mathbb{E}[d(X,X')] - \mathbb{E}[d(Y,Y')], \tag{1.7}$$

where $X, X' \overset{\text{i.i.d.}}{\sim} P$ and $Y, Y' \overset{\text{i.i.d.}}{\sim} Q$.

Proposition 1.9. Let $(\mathcal{X}, d)$ be a semimetric space. Suppose P and Q are two probability measure on $\mathcal{X}$ with finite first moments with respect to d. Then

(i) $D_{e,d}(P,Q) \geq 0$ for all such P and Q if and only if d is conditionally negative definite.

(ii) Furthermore, we have $D_{e,d}(P,Q) = 0 \ \Leftrightarrow \ P = Q$ if and only if d is strictly conditionally negative definite.

Now we continue our discussion in $\mathbb{R}^d$. We consider a translation invariant semimetric. Let $\Phi(\cdot - \cdot)$ be a semimetric in $\mathbb{R}^d$, where Φ is a nonnegative even function on $\mathbb{R}^d$, and we also say Φ is conditionally negative definite if $\Phi(\cdot - \cdot)$ is conditionally negative definite.

Definition 1.10 (Schwartz space). The Schwartz space $\mathcal{S}(\mathbb{R}^d)$ is defined as

$$\mathcal{S}(\mathbb{R}^d) = \left\{\gamma \in C^\infty(\mathbb{R}^d) : \forall \alpha, \beta \in \mathbb{N}_0^d, \ \sup_{x\in\mathbb{R}^d} \left|x^\alpha (D^\beta \gamma)(x)\right| < \infty\right\}, \tag{1.10}$$

where $x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$ and $D^\beta := \partial^{|\beta|}/\partial x_1^{\beta_1}\partial x_2^{\beta_2}\cdots\partial x_d^{\beta_d}$, $|\beta| = \beta_1 + \cdots + \beta_d$. Moreover, for $m \in \mathbb{N}$, we define

$$\mathcal{S}_m(\mathbb{R}^d) = \left\{\gamma \in \mathcal{S}(\mathbb{R}^d) : \gamma(x) = \mathcal{O}(\|x\|_2^m) \text{ for } \|x\|_2 \to 0\right\}. \tag{1.11}$$

Now we introduce the concept of generalized Fourier transform, which is useful in the analysis of conditionally negative definite functions. We say a function f is slowly increasing if it does not grow faster than a polynomial, that is, there exists $m \in \mathbb{N}_0$ such that $f(x) = \mathcal{O}(\|x\|_2^m)$ for $\|x\|_2 \to \infty$. The definition of generalized Fourier transform is presented below.

Definition 1.11 (Generalized Fourier transform). Suppose that $\Phi : \mathbb{R}^d \to \mathbb{C}$ is continuous and slowly increasing. A measurable function $\widehat{\Phi} \in L_2^{\mathrm{loc}}(\mathbb{R}^d \backslash \{0\})$ is said to be the generalized Fourier transform of Φ if there exists $m \in \mathbb{N}_0$ such that for all $\gamma \in \mathcal{S}_{2m}$,

$$\int_{\mathbb{R}^d} \Phi(x)\widehat{\gamma}(x)\mathrm{d}x = \int_{\mathbb{R}^d} \widehat{\Phi}(\omega)\gamma(\omega)\mathrm{d}\omega. \tag{1.12}$$

Here $\widehat{\gamma}(\omega) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} \gamma(x)\mathrm{e}^{-i\langle\omega,x\rangle}\mathrm{d}x$ is the Fourier transform of γ. The integer m is called the order of $\widehat{\Phi}$.

Theorem 1.12 (Bochner's characterization of conditionally negative definite functions). Suppose $\Phi : \mathbb{R}^d \to \mathbb{C}$ is continuous, slowly increasing, and possesses a generalized Fourier transform $\widehat{\Phi}$ of order 1, which is continuous on $\mathbb{R}^d \backslash \{0\}$. Then Φ is conditionally negative definite if $\widehat{\Phi}$ is negative and non-vanishing.

Some examples of conditionally negative function Φ are given in Corollary 1.13. Moreover, we can define the generalized energy distance:

$$D_{e,\Phi} =:= 2\mathbb{E}[\Phi(X-Y)] - \mathbb{E}[\Phi(X-X')] - \mathbb{E}[\Phi(Y-Y')], \tag{1.13}$$

where $X, X' \overset{\text{i.i.d.}}{\sim} P$ and $Y, Y' \overset{\text{i.i.d.}}{\sim} Q$.

Corollary 1.13. The following $\Phi : \mathbb{R} \to \mathbb{R}$ are conditionally negative definite functions on $\mathbb{R}^d$:

(i) $\Phi(x) = (c^2 + \|x\|_2^2)^\beta,\ 0 < \beta < 1$, with

$$\widehat{\Phi}(\omega) = -\frac{\beta 2^{1+\beta}}{\Gamma(1-\beta)} \left(\frac{\|\omega\|_2}{c}\right)^{-\beta-\frac{d}{2}} K_{\beta+\frac{d}{2}}(c\|\omega\|_2), \tag{1.14}$$

where $K_{\beta+d/2}$ is the modified Bessel function of the third kind of order $\beta + \frac{d}{2}$.

(ii) $\Phi(x) = \|x\|_2^\beta,\ 0 < \beta < 2$, with

$$\widehat{\Phi}(\omega) = -\frac{\beta 2^{\beta+\frac{d}{2}-1}\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma(1-\beta/2)} \|\omega\|_2^{-\beta-d}. \tag{1.15}$$

It can be seen that the classical energy distance $D_e = D_{e,\|\cdot\|_2}$ is a special case of the generalized definition $D_{e,\Phi}$, with the conditionally negative definite function $\Phi(x) = \|x\|_2$. Similar to Proposition 1.6, we have the following representation of generalized energy distance.

Proposition 1.14. Let $\Phi : \mathbb{R}^d \to [0, +\infty)$ be a conditionally negative definite semimetric on $\mathbb{R}^d$. Suppose P and Q are two probability measures on $\mathbb{R}^d$ such that $P, Q \in \mathcal{M}_\Phi^1(\mathbb{R}^d)$. Denote their characteristic functions by $\widehat{\mu}_P(\omega) = \int \mathrm{e}^{i\langle\omega,x\rangle}\mathrm{d}P(x)$ and $\widehat{\mu}_Q(\omega) = \int \mathrm{e}^{i\langle\omega,x\rangle}\mathrm{d}Q(x)$. Then the generalized energy distance between P and Q with respect to Φ admits the following representation:

$$D_{e,\Phi}(P,Q) = -(2\pi)^{-d/2} \int \widehat{\Phi}(\omega) \left|\widehat{\mu}_P(\omega) - \widehat{\mu}_Q(\omega)\right|^2 \mathrm{d}\omega, \tag{1.16}$$

where $\widehat{\Phi}$ is the generalized Fourier transform of Φ.

In fact, (1.16) can be viewed as a special case of (1.12) by setting $\gamma = |\widehat{\mu}_P - \widehat{\mu}_Q|^2$. Similarly, (1.4) is a special case of (1.16).

2 Maximum Mean Discrepancy

2.1 Kernel Embedding and Maximum Mean Discrepancy

Definition 2.1 (Reproducing kernel Hilbert space, RKHS). Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{X}$. Then $\mathcal{H}$ is a RKHS if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

(i) $k(\cdot, x) \in \mathcal{H}\ \forall x \in \mathcal{X}$, and

(ii) $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)\ \forall f \in \mathcal{H}$ and $x \in \mathcal{X}$.

Moreover, k is called the reproducing kernel of $\mathcal{H}$.

Based on the definition of RKHS, we introduce the definition of kernel embedding.

Definition 2.2 (Kernel embedding). Let $\mathcal{H}$ be a RKHS of real-valued functions on $\mathcal{X}$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let ν be a signed measure on $\mathcal{X}$. The kernel embedding of ν into $\mathcal{H}$ is a real-valued function $\mu_k(\nu) \in \mathcal{H}$ such that $\int f(x)\mathrm{d}\nu(x) = \langle f, \mu_k(\nu)\rangle_{\mathcal{H}}\ \forall f \in \mathcal{H}$.

It can be verified that $\mu_k(\nu)$ is well-defined if $\nu \in \mathcal{M}_k^{1/2}(\mathcal{X})$. Using the reproducing property, we have for any $y \in \mathcal{X}$ that

$$\mu_k(\nu)(y) = \langle \mu_k(\nu), k(\cdot, y)\rangle_{\mathcal{H}} = \int k(x, y)\mathrm{d}\nu(x). \tag{2.1}$$

Then the kernel embedding of ν can be alternatively defined by Bochner's integral $\mu_k(\nu) = \int k(\cdot, x)\mathrm{d}\nu(x)$. To ensure the existence of kernel embeddings, we need to define the class of measures that have finite θ-moments with respect to kernel k. Formally, define

$$\mathcal{M}_k^{\theta}(\mathcal{X}) = \left\{\nu \text{ is a finite signed measure on } \mathcal{X} : \int k^{\theta}(x, x)\mathrm{d}|\nu|(x) < \infty\right\}. \tag{2.2}$$

Definition 2.3 (Maximum mean discrepancy, MMD). Let $\mathcal{H}$ be a RKHS of real-valued functions on $\mathcal{X}$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and let $P, Q \in \mathcal{M}_k^{1/2}(\mathcal{X})$ be two probability measures on $\mathcal{X}$. The maximum mean discrepancy (MMD) γ_k between P and Q is defined as

$$\gamma_k(P, Q) := \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}}. \tag{2.3}$$

Using the Bochner's integral, the squared MMD can be represented as

$$\begin{aligned}
\gamma_k^2(P, Q) &= \left\|\int k(\cdot, x)\mathrm{d}P(x) - \int k(\cdot, x)\mathrm{d}Q(x)\right\|_{\mathcal{H}}^2 \\
&= \int\int k(x, y)\mathrm{d}[P - Q](x)\mathrm{d}[P - Q](y) \\
&= \mathbb{E}[k(X, X')] + \mathbb{E}[k(Y, Y')] - 2\mathbb{E}[k(X, Y)],
\end{aligned} \tag{2.4}$$

where $X, X' \overset{\text{i.i.d.}}{\sim} P$ and $Y, Y' \overset{\text{i.i.d.}}{\sim} Q$.

Proposition 2.4. Under the condition in Definition 2.3, we have

$$\gamma_k(P, Q) = \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \le 1} \int f\mathrm{d}[P - Q]. \tag{2.5}$$

This equation share a structure with the variational representation of total variation. Furthermore, we can derive a Koksma-Hlawka-like bound for the integration error. Suppose $f \in \mathcal{H}$, then we have

$$I(f; P, Q) := \left| \int f \mathrm{d}[P - Q] \right| \leq \gamma_k(P, Q) \|f\|_{\mathcal{H}}. \tag{2.6}$$

2.2 Equivalence of MMD and Energy Distance

Lemma 2.5 (Distance-induced kernels). Let $(\mathcal{X}, d)$ be a semimetric space, and let $x_0 \in \mathcal{X}$. The kernel induced by d and centered at x_0 is defined as

$$k(x, x') = \frac{1}{2} \left[d(x, x_0) + d(x', x_0) - d(x, x') \right]. \tag{2.7}$$

Moreover, k is positive definite if and only if d is conditionally negative definite.

Lemma 2.5 establishes a connection between conditionally negative definite semimetric and positive definite kernel. By applying Moore-Aronszajn theorem, given a conditionally negative definite semimetric d, we can construct a RKHS with reproducing kernel defined in (2.7). Note that the constructed RKHS is not unique, since we can choose different centers $x_0 \in \mathcal{X}$. In general, the following properties hold.

Proposition 2.6. Let $(\mathcal{X}, d)$ be a semimetric space where d is conditionally negative definite. Let k be the kernel induced by d and centered at $x_0 \in \mathcal{X}$. The following statements are true:

(i) k is non-degenerate. That is, the Aronszajn map $x \mapsto k(\cdot, x)$ is injective;

(ii) $d(x, x') = k(x, x) + k(x', x') - 2k(x, x') = \|k(\cdot, x) - k(\cdot, x')\|_{\mathcal{H}}^2$.

The following proposition is an immediate corollary of Proposition 1.3.

Proposition 2.7. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric and positive definite kernel on $\mathcal{X}$. If k is non-degenerate, that is, the Aronszajn map $x \mapsto k(\cdot, x)$ is injective, then

$$d(x, x') = k(x, x) + k(x', x') - 2k(x, x') \tag{2.8}$$

defines a conditionally negative definite semimetric on $\mathcal{X}$. The conditionally negative definite function d is said to be generated by kernel k.

The existence of kernel embedding through a semimetric is discussed by Sejdinovic et al. (2013), in which a detailed proof of the following proposition can be find.

Proposition 2.8. Let $(\mathcal{X}, d)$ be a semimatric space where d is conditionally negative definite. Let k be a positive definite kernel that generates d. Then for any $n \in \mathbb{N}$, $\mathcal{M}_d^{n/2}(\mathcal{X}) = \mathcal{M}_k^{n/2}(\mathcal{X})$.

Based on the previous discussions, we immediately have the following theorem.

Theorem 2.9. Let $(\mathcal{X}, d)$ be a semimatric space where d is conditionally negative definite. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be any kernel that generates d. Then for any probability measures $P, Q \in \mathcal{M}_d^1(\mathcal{X})$, it holds

$$D_{e,d}(P, Q) = 2\gamma_k^2(P, Q). \tag{2.9}$$

2.3 Universal Kernels

In this subsection we introduce the universal kernels. We first investigate some properties of bounded kernels. Note that a valid kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ must be positive definite, we have

$$|k(x, x')| \leq \sqrt{k(x, x)k(x', x')}. \tag{2.10}$$

Hence k is bounded if and only if

$$\|k\|_\infty := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)} < \infty. \tag{2.11}$$

The following lemma provides another important characterization of bounded kernels.

Proposition 2.10 (Characterization of bounded kernels). Let $\mathcal{X}$ be a set and k be a kernel on $\mathcal{X}$ with corresponding RKHS $\mathcal{H}$. Then k is bounded if and only if every $f \in \mathcal{H}$ is bounded. Furthermore, in this case the inclusion map $\iota : \mathcal{H} \to L^\infty(\mathcal{X})$ is well-defined and continuous, and $\|\iota\| = \|k\|_\infty$.

Furthermore, we may be interested in the continuity of functions in a RKHS. A characterization of RKHS's of bounded and continuous functions, denoted by $C_b(\mathcal{X}) := C(\mathcal{X}) \cap L^\infty(\mathcal{X})$, is presented as follows.

Proposition 2.11. Let $\mathcal{X}$ be topological space and k be a kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$. Then k is bounded and separately continuous if and only if every $f \in \mathcal{H}$ is a bounded and continuous function. In this case, the inclusion map $\iota : \mathcal{H} \to C_b(\mathcal{X})$ is well-defined and continuous and we have $\|\iota\| = \|k\|_\infty$.

Now we are ready to introduce the universal kernel.

Definition 2.12 (Universal kernel). Let k be a continuous kernel on a compact metric space $\mathcal{X}$, and let $\mathcal{H}$ be the corresponding RKHS. Then k is said to be universal, if $\mathcal{H}$ is dense in $C(\mathcal{X})$ with respect to $\|\cdot\|_\infty$, i.e. for every $g \in C(\mathcal{X})$ and all $\epsilon > 0$, there exists an $f \in \mathcal{H}$ such that

$$\|f - g\|_\infty < \epsilon. \tag{2.12}$$

Proposition 2.13. Let $\mathcal{X}$ be a compact metric space and k be a universal kernel on $\mathcal{X}$ with corresponding RKHS $\mathcal{H}$. The following statements are true:

(i) k separates all compact sets in $\mathcal{X}$, that is, for all compact disjoint sets $\mathcal{A}, \mathcal{B} \subset \mathcal{X}$, there exists $f \in \mathcal{H}$ such that $f(x) > 0\ \forall x \in \mathcal{A}$ and $f(x) < 0\ \forall x \in \mathcal{B}$;

(ii) There exists $L > 0$ such that $k(x, x) \geq L$ for all $x \in \mathcal{X}$.

In some practical scenarios such as hypothesis testing, we may be concerned about whether the kernel embedding μ_k is injective, which determines the ability of MMD to distinguish two different probability measures. For universal kernels, we have the following result.

Theorem 2.14. Let $\mathcal{X}$ be a compact metric space and k be a universal kernel on $\mathcal{X}$. Then for two probability measures on $\mathcal{X}$ with their kernel embeddings exist, $\gamma_k(P, Q) = 0$ if and only if $P = Q$.

3 f-Divergence

In this section, we introduce the f-divergence between probability measures over a measurable space $(\mathcal{X}, \Sigma)$. All f-divergences quantify the difference between a pair of measures or distributions, each with different operational meaning.

3.1 f-Divergence

Definition 3.1 (f-divergence). Let P and Q be two probability measures on a measurable space $(\mathcal{X}, \Sigma)$ with $Q \ll P$. Then for any convex function $f : [0, +\infty) \to (-\infty, +\infty]$ such that (i) $f(1) = 0$, (ii) f is strictly convex at 1, and (iii) f is finite except possibly at 0, the f-divergence of Q with respect to P is defined as

$$D_f(Q\|P) := \int f\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)\mathrm{d}P, \tag{3.1}$$

where the notation $\frac{\mathrm{d}Q}{\mathrm{d}P}$ stands for the Radon-Nikodym derivative of Q with respect to P.

The above definition is not convenient when it comes to calculation. In practice, we often use the following two forms of f-divergence:

- When $\mathcal{X}$ is discrete, P and Q are probability mass functions:

$$D_f(Q\|P) = \sum_{x\in\mathcal{X}} f\left(\frac{Q(x)}{P(x)}\right)P(x). \tag{3.2}$$

- When P and Q are characterized by density functions p and q (i.e. their Radon-Nikodym derivatives with respect to the Lebesgue measure), respectively, then

$$D_f(q\|p) = \int f\left(\frac{q(x)}{p(x)}\right)p(x)\mathrm{d}x. \tag{3.3}$$

Definition 3.2 (Examples of f-divergence). The following are some commonly used f-divergences:

- **Total variation distance.** $f(x) = \frac{1}{2}|x-1|$:

$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\int\left|\frac{\mathrm{d}Q}{\mathrm{d}P} - 1\right|\mathrm{d}P = \frac{1}{2}\int|\mathrm{d}Q - \mathrm{d}P|. \tag{3.4}$$

  Note that $d_{\mathrm{TV}}(P,Q) = d_{\mathrm{TV}}(Q,P)$.

- **Kullback-Leibler divergence.** $f(x) = x\log x$:

$$D_{\mathrm{KL}}(Q\|P) = \int\log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)\mathrm{d}Q. \tag{3.5}$$

- **Squared Hellinger distance.** $f(x) = (1-\sqrt{x})^2$:

$$H^2(P,Q) = \int\left(1 - \sqrt{\frac{\mathrm{d}Q}{\mathrm{d}P}}\right)^2\mathrm{d}P = \int\left(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q}\right)^2. \tag{3.6}$$

  Note that $H^2(P,Q) = H^2(Q,P)$.

- **Pearson χ^2-divergence.** $f(x) = x^2 - 1$:

$$\chi^2(Q\|P) = \int\frac{\mathrm{d}Q^2}{\mathrm{d}P} - 1. \tag{3.7}$$

- **Jensen-Shannon divergence.** $f(x) = \frac{x}{2}\log x - \frac{1+x}{2}\log\left(\frac{1+x}{2}\right)$:

$$d_{\mathrm{JS}}(P,Q) = \frac{1}{2}D_{\mathrm{KL}}(P\|M) + \frac{1}{2}D_{\mathrm{KL}}(Q\|M), \tag{3.8}$$

  where $M = \frac{1}{2}P + \frac{1}{2}Q$. It is also known as the symmetrized Kullback-Leibler divergence.

3.2 Variational Representation

Before we introduce the variational representation of the f-divergence, let's review the Fenchel conjugate.

Definition 3.3 (Fenchel conjugate). Let $\mathcal{X}$ be a real Hilbert space equipped with an inner product $\langle \cdot, \cdot \rangle$, and let $f : \mathcal{X} \to (-\infty, +\infty]$ be a proper function, that is, $\mathrm{dom}(f) := \{x \in \mathcal{X} : f(x) \in \mathbb{R}\} \neq \emptyset$. The Fenchel conjugate of f is defined as

$$f^*(t) = \sup_{x \in \mathcal{X}} \{\langle x, t \rangle - f(x)\}, \ t \in \mathcal{X}. \tag{3.9}$$

It can be seen that f^* is the pointwise supremum of a collection of affine functions, hence f^* is convex, regardless of f is convex or not. Moreover, it can be shown that the duality $(f^*)^* = f$ holds if f is convex and lower semicontinuous.

Below is an immediate consequence of Definition 3.3.

Proposition 3.4 (Fenchel-Young inequality). $\forall x, t \in \mathcal{X}$,

$$f(x) + f^*(t) \geq \langle x, t \rangle. \tag{3.10}$$

Recall that in Definition 3.1, f is defined on $[0, +\infty)$. We complete f by redefining $f(x) = \infty$ for $x \in \mathbb{R}$ with $x < 0$. This operation preserves the convexity of f. Moreover, the Fenchel conjugate of $f : \mathbb{R} \to (-\infty, +\infty]$ is well defined: $f^*(t) = \sup_{x \in \mathbb{R}} \{tx - f(x)\}, \ t \in \mathbb{R}$.

The f-divergence admits the following variational representation.

Lemma 3.5 (Variational representation of f-divergence). Denote $\mathcal{G}$ by the class of measurable functions on $(\mathcal{X}, \Sigma)$. Then the f-divergence can be represented as

$$D_f(Q \| P) = \sup_{g \in \mathcal{G}} \left\{ \int g \mathrm{d}Q - \int (f^* \circ g) \mathrm{d}P \right\}. \tag{3.11}$$

Where f is the Fenchel conjugate of f. If f is differentiable, then the supremum is reached at $\widetilde{g} = f'(\mathrm{d}Q/\mathrm{d}P)$.

Proposition 3.6 (Variational representations of f-divergences in Definition 3.2).

- **Total variation distance**. $f^*(t) = \begin{cases} t, \ |t| \leq 1/2 \\ \infty, \ |t| > 1/2 \end{cases}$:

$$d_{\mathrm{TV}}(P, Q) = \frac{1}{2} \sup_{\|g\|_\infty \leq 1} \int g \mathrm{d}[P - Q]. \tag{3.12}$$

- **Kullback-Leibler divergence.** $f^*(t) = \mathrm{e}^{t-1}$:

$$D_{\mathrm{KL}}(Q \| P) = 1 + \sup_{g : \mathcal{X} \to \mathbb{R}} \left\{ \int g(x) \mathrm{d}Q(x) - \int \exp[g(x)] \mathrm{d}P(x) \right\}. \tag{3.13}$$

- **Squared Hellinger distance.** $f^*(t) = \begin{cases} \frac{t}{1-t}, \ t < 1 \\ \infty, \ t \geq 1 \end{cases}$:

$$H^2(P, Q) = 2 - \inf_{g > 0} \left\{ \int g \mathrm{d}Q + \int \frac{1}{g} \mathrm{d}P \right\}. \tag{3.14}$$

- **Pearson χ^2-divergence.** $f^*(t) = \frac{1}{4}t^2 + 1$:

$$\chi^2(Q\|P) = \sup_{g:\mathcal{X}\to\mathbb{R}} \left\{ \int g\mathrm{d}Q - \frac{1}{4}\int g^2\mathrm{d}P \right\} - 1. \tag{3.15}$$

  Let $g = a + bh$, and solve (3.15) with respect to a, b, we obtain a more symmetric version which is directly related to the bias-variance tradeoff:

$$\chi^2(Q\|P) = \sup_{h:\mathcal{X}\to\mathbb{R}} \frac{\left(\int h\mathrm{d}[Q-P]\right)^2}{\int h^2\mathrm{d}P - \left(\int h\mathrm{d}P\right)^2}. \tag{3.16}$$

- **Jensen-Shannon divergence.** $f^*(t) = \begin{cases} -\frac{1}{2}\log(2-\mathrm{e}^{2t}),\ t < \frac{1}{2}\log 2 \\ \infty,\ t \geq \frac{1}{2}\log 2. \end{cases}$:

$$d_{\mathrm{JS}}(P,Q) = \frac{1}{2}\sup_{\|f\|_\infty < 1} \left\{ \int \log(1+h)\mathrm{d}Q + \int \log(1-h)\mathrm{d}P \right\}. \tag{3.17}$$

3.3 Inequality between f-divergences

Theorem 3.7 (Pinsker's inequality). If P and Q are two probability measures on a measurable space $(\mathcal{X}, \Sigma)$ with $Q \ll P$, then

$$d_{\mathrm{TV}}(P,Q) \leq \sqrt{\frac{1}{2}D_{\mathrm{KL}}(Q\|P)}. \tag{3.18}$$

Proposition 3.8 (The Bretagnolle-Huber bound). If P and Q are two probability measures on a measurable space $(\mathcal{X}, \Sigma)$ with $Q \ll P$, then

$$d_{\mathrm{TV}}(P,Q) \leq \sqrt{1-\exp\left(-D_{\mathrm{KL}}(Q\|P)\right)} \leq 1 - \frac{1}{2}\exp\left(-D_{\mathrm{KL}}(Q\|P)\right). \tag{3.19}$$

Proposition 3.9. If P and Q are two probability measures on a measurable space $(\mathcal{X}, \Sigma)$, then

$$d_{\mathrm{JS}}(P,Q) \leq d_{\mathrm{TV}}(P,Q). \tag{3.20}$$

3.4 f-Divergence and MMD

Theorem 3.10 (Total variation and MMD). Let P and Q be two probability measures on measurable $(\mathcal{X}, \Sigma)$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded kernel. Then

$$\gamma_k(P,Q) \leq 2\|k\|_\infty d_{\mathrm{TV}}(P,Q). \tag{3.21}$$

4 Proofs

4.1 Proof of Proposition 1.3

Proof. **"If" part:** $\forall n \in \mathbb{N},\ x_1, \cdots, x_n \in \mathcal{X}$ and $\alpha_1, \cdots, \alpha_n \in \mathbb{R}$ with $\sum_{j=1}^n \alpha_j = 0$,

$$\begin{aligned}
\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j d(x_i, x_j) &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \langle \phi(x_i) - \phi(x_j), \phi(x_i) - \phi(x_j) \rangle_{\mathcal{H}} \\
&= \sum_{i=1}^n \alpha_i \sum_{j=1}^n \alpha_j \|\phi(x_j)\|_2^2 + \sum_{i=1}^n \alpha_i \|\phi(x_i)\|_2^2 \sum_{j=1}^n \alpha_j - 2 \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} \\
&= -2 \left\langle \sum_{i=1}^n \alpha_i \phi(x_i), \sum_{j=1}^n \alpha_j \phi(x_j) \right\rangle_{\mathcal{H}} = \left\| \sum_{j=1}^n \alpha_j \phi(x_j) \right\|_{\mathcal{H}}^2 \leq 0.
\end{aligned}$$

"Only if" part: Suppose d is conditionally negative definite on $\mathcal{X}$, i.e., $\forall n \in \mathbb{N},\ x_1, \cdots, x_n \in \mathcal{X}$ and $\alpha_1, \cdots, \alpha_n \in \mathbb{R}$ with $\sum_{j=1}^n \alpha_j = 0$, d satisfies (1.1). We choose an arbitrary $x_0 \in \mathcal{X}$, and define a map $\phi : x \mapsto \frac{1}{2} \left[d(\cdot, x_0) + d(x, x_0) - d(\cdot, x) \right]$.

We construct a vector space $\mathcal{H}_0 := \operatorname{span}\{\phi(x) : x \in \mathcal{X}\}$ and a bilinear form $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$, which is defined as

$$\langle f, g \rangle_{\mathcal{H}_0} = \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^n f_i g_j \left[d(x_i, x_0) + d(y_j, x_0) - d(x_i, y_j) \right], \quad f := \sum_{i=1}^m f_i \phi(x_i),\ g := \sum_{j=1}^n g_j \phi(y_j),$$

where $f_1, \cdots, f_m, g_1, \cdots, g_n \in \mathbb{R}, x_1, \cdots, x_m, y_1, \cdots, y_n \in \mathcal{X}$.

Now we verify that $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is a valid inner product on $\mathcal{H}_0$. The linearity of $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ holds by definition, and the symmetry of $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ follows from the symmetry of d. It remains to show the nonnegativity. Denote $g_0 = -\sum_{j=1}^n g_j$ and $y_0 = x_0$, we have:

$$\begin{aligned}
\langle g, g \rangle_{\mathcal{H}_0} &= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n g_i g_j \left[d(y_i, x_0) + d(y_j, x_0) - d(y_i, y_j) \right] \\
&= -\frac{1}{2} \sum_{i=0}^n \sum_{j=0}^n g_i g_j d(y_i, y_j) \geq 0,
\end{aligned}$$

where the inequality follows from (1.1). Hence $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is a semi-inner product. Specifically, we have

$$\|\phi(x) - \phi(x')\|_{\mathcal{H}_0}^2 = d(x, x').$$

Furthermore, in order to show that $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is an inner product, we need to argue that $\langle g, g \rangle_{\mathcal{H}_0}$ implies $g \equiv 0$. We first show that $g \equiv 0$ if and only if $\langle f, g \rangle_{\mathcal{H}_0} = 0\ \forall f \in \mathcal{H}_0$. Since $\mathcal{H}_0 = \operatorname{span}\{\phi(x) : x \in \mathcal{X}\}$, it suffices to show that $g \equiv 0$ if and only if $\langle \phi(x), g \rangle_{\mathcal{H}_0} = 0\ \forall x \in \mathcal{X}$:

$$\langle \phi(x), g \rangle_{\mathcal{H}_0} = \frac{1}{2} \sum_{j=1}^n g_j \left[d(x, x_0) + d(y_j, x_0) - d(x, y_j) \right] = g(x).$$

Suppose $\langle g, g \rangle_{\mathcal{H}_0} = 0$, then $\forall t \in \mathbb{R}$ and $f \in \mathcal{H}_0$, we have

$$0 \leq \langle g - tf, g - tf \rangle_{\mathcal{H}_0} = t^2 \|f\|_{\mathcal{H}_0} - 2t \langle g, f \rangle_{\mathcal{H}_0},$$

where the discrinant of RHS is $\Delta = 4 |\langle g, f \rangle_{\mathcal{H}_0}|^2 \leq 0$. Hence $\langle g, f \rangle_{\mathcal{H}_0} = 0\ \forall f \in \mathcal{H}_0$, and $g \equiv 0$.

Now we complete the space $\mathcal{H}_0$ by taking equivalence classes of Cauchy sequences from $\mathcal{H}_0$. Then we obtain a Hilbert space $\mathcal{H}$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$.

Let $\{f_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence in $\mathcal{H}_0$. By Cauchy-Schwarz inequality,

$$|f_n(x) - f_m(x)| \leq \|f_n - f_m\|_{\mathcal{H}_0}\|\phi(x)\|_{\mathcal{H}_0},$$

hence the sequence is pointwisely Cauchy in $\mathbb{R}$, and we define $f(x) = \lim_{n\to\infty} f_n(x) \in \mathcal{H}$. Similarly, we denote the limit of Cauchy sequence $\{g_n\}_{n\in\mathbb{N}}$ by g. For the f and g defined as above, let

$$\langle f, g\rangle_{\mathcal{H}} := \lim_{n\to\infty} \langle f_n, g_n\rangle_{\mathcal{H}_0}.$$

Then $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ defines an inner product on $\mathcal{H}$.

It remains to show that $\phi : \mathcal{X} \to \mathcal{H}$ is an injective map. For any $x, x' \in \mathcal{X}$,

$$\begin{aligned}
\phi(x) = \phi(y) \;&\Rightarrow\; d(x, x_0) - d(\cdot, x) = d(x', x_0) - d(\cdot, x') \\
&\Rightarrow \begin{cases} d(x, x_0) = d(x', x_0) - d(x, x'), \\ d(x, x_0) - d(x', x) = d(x', x_0), \end{cases} \\
&\Rightarrow\; d(x', x) = 0 \Rightarrow x = x'.
\end{aligned}$$

Thus we conclude our proof by finding such a Hilbert space $\mathcal{H}$ and an injective map $\phi : \mathcal{X} \to \mathcal{H}$ that satisfy the condition in Proposition 1.3. □

4.2 Proof of Proposition 1.9

Proof. For $P, Q \in \mathcal{M}_d^1(\mathcal{X})$, we have

$$D_{e,d}(P, Q) = -\int\int d(x, y)\mathrm{d}[P - Q](x)\mathrm{d}[P - Q](y)$$

(i) **"Only if" part:** Suppose $D_{e,d}(P, Q) \geq 0$ for all $P, Q \in \mathcal{M}_d^1(\mathcal{X})$.

Then $\forall n \in \mathbb{N}$, $x_1, \cdots, x_n \in \mathcal{X}$ and $\alpha_1, \cdots, \alpha_n \in \mathbb{R}$ such that $\sum_{j=1}^n \alpha_j = 0$, we need to show that the inequality (1.1) holds. Without loss of generality, let $x_1, \cdots, x_n$ be mutually distinct, and let $\alpha_1, \cdots, \alpha_n$ be nonzero. We use the notations $[x]_+ = \max\{x, 0\}$ and $[x]_- = \max\{-x, 0\}$ standing for the positive and negative part of a real number. Since $\sum_{j=1}^n \alpha_j = 0$, let $A = \sum_{j=1}^n [\alpha_j]_+ = \sum_{j=1}^n [\alpha_j]_- > 0$. Then inequality (1.1) immediately holds by choosing the following two discrete measures:

$$P(x) = \begin{cases} [\alpha_j]_+/A, \ \exists x = x_j, \\ 0, \ x \in \mathcal{X}\backslash\{x_1, \cdots, x_n\}, \end{cases} \qquad Q(x) = \begin{cases} [\alpha_j]_-/A, \ \exists x = x_j, \\ 0, \ x \in \mathcal{X}\backslash\{x_1, \cdots, x_n\}. \end{cases}$$

"If" part: Suppose d is conditionally negative definite. Then $\forall P, Q \in \mathcal{M}_d^1(\mathcal{X})$, we need to show that $D_e(P, Q) \geq 0$. We can finish the proof by the simple approximation theorem.

(ii) is a immediate corollary of (i).

□

4.3 Proof of Proposition 1.14

Proof. Let $\gamma(\omega) = |\widehat{\mu}_P(\omega) - \widehat{\mu}_Q(\omega)|^2$, we first show that $\gamma \in \mathcal{S}_2$.

... □

4.4 Sufficient Condition for the Existence of Kernel Embedding

For a finite signed measure ν on $\mathcal{X}$, the existence of kernel embedding $\mu_k(\nu)$ in Definition 2.2 can be ensured by some regularity conditions. To see this, we define a linear functional $T_\nu : f \mapsto \int f(x)\mathrm{d}\nu(x)$. From the Riesz representation theorem, if T_ν is continuous, then we can find a unique $\mu_k(\nu) \in \mathcal{H}$ such that $T_\nu f = \langle f, \mu_k(\nu)\rangle_{\mathcal{H}}$, which is the Riesz representation of T_ν. This is equivalent to find ν such that the norm of functional T_ν is bounded:

$$\|T_\nu\|_{\mathcal{H}^*} = \sup_{f\in\mathcal{H}\setminus\{0\}} \frac{|T_\nu f|}{\|f\|_{\mathcal{H}}} < \infty.$$

For the numerator, we have

$$\begin{aligned}
|T_\nu f| = \left|\int f(x)\mathrm{d}\nu(x)\right| &= \left|\int \langle f, k(\cdot,x)\rangle_{\mathcal{H}}\mathrm{d}\nu(x)\right| \\
&= \left|\left\langle f, \int k(\cdot,x)\mathrm{d}\nu(x)\right\rangle_{\mathcal{H}}\right| \\
&\le \|f\|_{\mathcal{H}} \left\|\int k(\cdot,x)\mathrm{d}\nu(x)\right\|_{\mathcal{H}} \\
&= \|f\|_{\mathcal{H}} \sqrt{\int\int k(x,y)\mathrm{d}\nu(x)\mathrm{d}\nu(y)}
\end{aligned}$$

It is seen that T_ν is continuous when the integral $\int\int k(x,y)\mathrm{d}\nu(x)\mathrm{d}\nu(y)$ is bounded. Some stronger assumptions are presented below.

Proposition A.1. (i) If $\sup_{x\in\mathcal{X}} k(x,x) < \infty$, then $\mu_k(\nu)$ is well-defined; (ii) If $\nu \in \mathcal{M}_k^{1/2}(\mathcal{X})$, then $\mu_k(\nu)$ is well-defined.

Here (i) holds since ν is a finite signed measure on $\mathcal{X}$. For (ii), note that k is positive definite, we have $|k(x,y)| \le \sqrt{k(x,x)k(y,y)}$. Then

$$\begin{aligned}
\int\int k(x,y)\mathrm{d}\nu(x)\mathrm{d}\nu(y) &\le \int\int |k(x,y)|\mathrm{d}|\nu(x)|\mathrm{d}|\nu(y)| \\
&\le \int\int k^{1/2}(x,x)k^{1/2}(y,y)\mathrm{d}|\nu|(x)\mathrm{d}|\nu|(y) \\
&= \left(\int k^{1/2}(x,x)\mathrm{d}|\nu|(x)\right)^2 < \infty,
\end{aligned}$$

where the last inequality holds by $\nu \in \mathcal{M}_k^{1/2}(\mathcal{X})$.

4.5 Proof of Proposition 2.4

Proof. For any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} = 1$, we have

$$\begin{aligned}
\int f\mathrm{d}[P-Q] &= \int \langle f, k(\cdot,x)\rangle_{\mathcal{H}}\mathrm{d}[P-Q](x) \\
&= \left\langle f, \int k(\cdot,x)\mathrm{d}[P-Q](x)\right\rangle_{\mathcal{H}} \\
&= \langle f, \mu_k(P) - \mu_k(Q)\rangle_{\mathcal{H}} \\
&\le \|f\|_{\mathcal{H}}\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}} =: \gamma_k(P,Q),
\end{aligned}$$

where the first equality follows from the reproducing property of k, the second from the continuity of inner product, the third by definition, and the inequality is Cauchy-Schwarz. Furthermore, the equality holds when

$$f = \frac{\mu_k(P) - \mu_k(Q)}{\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}}},$$

where the supremum in (2.5) is reached. $\square$

4.6 Proof of Lemma 2.5

Proof. $\forall n \in \mathbb{N},\ x_1, \cdots, x_n \in \mathcal{X}$ and $c_1, \cdots, c_n \in \mathbb{R}$, denote $c_0 = -\sum_{j=1}^n c_j$. Then

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^n c_i c_j k(x_i, x_j) &= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n c_i c_j \left[d(x_i, x_0) + d(x_j, x_0) - d(x_i, x_j)\right] \\ &= -\frac{1}{2} \sum_{i=0}^n \sum_{j=1}^n c_i c_j d(x_i, x_j).\end{aligned}$$

Therefore k is positive definite if and only if d is conditionally negative definite. $\square$

4.7 Proof of Theorem 2.9

Proof. For any P and Q that have finite first moments with respect to d,

$$\begin{aligned}D_{e,d}(P, Q) &:= 2\int d(x, y)\mathrm{d}[P \times Q](x, y) - \int d(x, x')\mathrm{d}[P \times P](x, x') - \int d(y, y')\mathrm{d}[Q \times Q](y, y') \\ &= -\int\int d(x, y)\mathrm{d}[P - Q](x)\mathrm{d}[P - Q](y) \\ &= -\int\int [k(x, x) + k(y, y) - 2k(x, y)]\mathrm{d}[P - Q](x)\mathrm{d}[P - Q](y) \\ &= 2\int\int k(x, y)\mathrm{d}[P - Q](x)\mathrm{d}[P - Q](y) =: 2\gamma_k^2(P, Q),\end{aligned}$$

where the last equality uses $\int \mathrm{d}[P - Q] = 0$. $\square$

4.8 Proof of Proposition 2.10

Proof. **"If" part:** Suppose every $f \in \mathcal{H}$ is bounded. Then the inclusion map $\iota : \mathcal{H} \to L^\infty(\mathcal{X})$ is well-defined. We fix a sequence $\{f_n\} \subset \mathcal{H}$ for which $\exists f \in \mathcal{H}$ and $g \in L^\infty(\mathcal{X})$ such that

$$\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}} = 0,\ \lim_{n\to\infty} \|\iota f_n - g\|_\infty = 0.$$

Both the two convergence implies pointwise convergence:

$$\begin{aligned}|f_n(x) - f(x)| &= |\langle f - f_n, k(\cdot, x)\rangle|_{\mathcal{H}} \leq \sqrt{k(x, x)}\|f - f_n\|_{\mathcal{H}} \to 0, \\ |f_n(x) - g(x)| &\leq \sup_{x\in\mathcal{X}} |\iota f_n(x) - g(x)| = \|\iota f_n(x) - g(x)\|_\infty \to 0.\end{aligned}$$

Then $f = g$. By the closed graph theorem, ι is continuous, and

$$k(x, x) \leq \|k(\cdot, x)\|_\infty \leq \|\iota\|\|k(\cdot, x)\|_{\mathcal{H}} \leq \|\iota\|\sqrt{k(x, x)}.$$

Then we have $\|k\|_\infty \leq \|\iota\|$.

"Only if" part: Suppose k is bounded. For any $f \in \mathcal{H}$, we have for all $x \in \mathcal{X}$ that

$$\begin{aligned} |f(x)| = |\langle f, k(\cdot, x)\rangle_{\mathcal{H}}| &\leq \|f\|_{\mathcal{H}} \|k(\cdot, x)\|_{\mathcal{H}} \\ &= \|f\|_{\mathcal{H}} \sqrt{k(x,x)} \leq \|f\|_{\mathcal{H}} \|k\|_{\infty}. \end{aligned}$$

Hence $\|f\|_\infty \leq \|f\|_{\mathcal{H}} \|k\|_\infty$, and the inclusion map $\iota : \mathcal{H} \to L^\infty(\mathcal{X})$ is well-defined. Moreover,

$$\|\iota\| = \sup_{f \in \mathcal{H}} \frac{\|f\|_\infty}{\|f\|_{\mathcal{H}}} \leq \|k\|_\infty.$$

The two inequalities imply $\|k\|_\infty = \|\iota\|$. $\square$

4.9 Proof of Proposition 2.11

Proof. **"If" part:** Suppose that every $f \in \mathcal{H}$ is bounded and continuous. Then $k(\cdot, x) : \mathcal{X} \to \mathbb{R}$ is continuous for every $x \in \mathcal{X}$. Moreover, the boundedness of k is ensured by Proposition 2.10.

"Only if" part: Suppose that k is bounded and separately continuous. Then the pre-Hilbert space $\mathcal{H}_0 = \mathrm{span}\{k(\cdot, x) : x \in \mathcal{X}\}$ only contains continuous functions. Since $\mathcal{H}$ is complete, and k is bounded, we can choose a sequence $\{f_n\} \subset \mathcal{H}_0$ for any f that satisfies

$$0 = \lim_{n \to \infty} \|f_n - f\|_{\mathcal{H}} \geq \frac{1}{\|k\|_\infty} \lim_{n \to \infty} \|f_n - f\|_\infty \geq 0.$$

Hence $\{f_n\}$ converges uniformly to f. Note that $\{f_n\}$ are continuous, $\forall \epsilon > 0$, there exists $\eta > 0$ such that $\sup_{x' : d(x,x') < \eta} |f_n(x) - f_n(x')| < \epsilon/3$. Moreover, there exists N such that for all $n > N$, $\|f_n - f\|_\infty < \epsilon/3$. Then

$$\sup_{x' : d(x,x') < \eta} |f(x) - f(x')| \leq \sup_{x' : d(x,x') < \eta} \{|f(x) - f_n(x)| + |f_n(x) - f_n(x')| + |f_n(x') - f(x')|\} \leq \epsilon.$$

Therefore f is continuous. The remaining part is similar to Proposition 2.10. $\square$

4.10 Proof of Proposition 2.13

Proof. Let $\mathcal{A}, \mathcal{B} \subset \mathcal{X}$ be disjoint compact subsets and d be the metric of $\mathcal{X}$. Then, $\forall x \in \mathcal{X}$, define

$$g(x) := \frac{d(x, \mathcal{A}) - d(x, \mathcal{B})}{d(x, \mathcal{A}) + d(x, \mathcal{B})},$$

where the distance function is defined as $d(x, \mathcal{C}) = \inf_{x' \in \mathcal{C}} d(x, x')$ for $x \in \mathcal{X}$ and $\mathcal{C} \subset \mathcal{X}$. We claim that $d(x, \mathcal{C})$ is continuous with respect to x. To see this, fix an arbitrary $\epsilon > 0$. Then $\forall x \in \mathcal{X}$, there exists $x' \in \mathcal{C}$ such that $d(x, x') < d(x, \mathcal{C}) + \epsilon$. For all $x, y \in \mathcal{X}$ with $d(x, y) < \epsilon$,

$$d(x, \mathcal{C}) > d(x, x') - \epsilon \geq d(y, x') - d(x, y) - \epsilon > d(y, \mathcal{C}) - 2\epsilon.$$

Similarly, we have $d(y, \mathcal{C}) > d(x, \mathcal{C}) - 2\epsilon$. Hence $|d(\mathcal{C}, y) - d(x, \mathcal{C})| < 2\epsilon$, and $d(\cdot, \mathcal{C})$ is continuous. Furthermore, $g(x)$ is continuous.

Moreover, note that $g(x) = 1 \ \forall x \in \mathcal{A}$, and $g(x) = -1 \ \forall x \in \mathcal{B}$. Since k is universal, there exists $h \in \mathcal{H}$ such that $|h(x) - g(x)| < \frac{1}{2}$ for all $x \in \mathcal{X}$. Then $h(x) > \frac{1}{2} \ \forall x \in \mathcal{A}$, and $h(x) < -\frac{1}{2} \ \forall x \in \mathcal{B}$. Hence the statement (i) holds.

For the statement (ii), we choose $g \equiv 1$. Then there exists $f \in \mathcal{H}$ such that $|f(x) - g(x)| < \frac{1}{2}$ for all $x \in \mathcal{X}$.

From the reproducing property, we have

$$0 < f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} \sqrt{k(x,x)}.$$

Then $k(x,x) > 0$ for all $x \in \mathcal{X}$. Note that $\mathcal{X}$ is compact and that k is continuous, there exists $x^* \in \mathcal{X}$ such that $0 < k(x^*, x^*) = \inf_{x\in\mathcal{X}} k(x,x) =: L$. $\square$

4.11 Proof of Theorem 2.14

Proof. It suffices to show that $\gamma_k(P,Q) = 0$ implies $P = Q$, since the converse holds whenever their kernel embedding exists. We first show that $\gamma_k(P,Q) = 0$ implies $\int f\mathrm{d}P = \int f\mathrm{d}Q$ for all $f \in C(\mathcal{X})$.

We argue this by contradiction. Suppose $\sup_{f\in C(\mathcal{X})} \left|\int f\mathrm{d}[P-Q]\right| = L > 0$, then there exists $\tilde{f} \in C(\mathcal{X})$ with $\left|\int \tilde{f}\mathrm{d}[P-Q]\right| > L/2$. From the universality of kernel k, the corresponding RKHS $\mathcal{H}$ is dense in $C(\mathcal{X})$ with respect to $\|\cdot\|_\infty$. Then there exists $h \in \mathcal{H}$ such that $\|h - \tilde{f}\|_\infty < L/8$, and

$$0 = \gamma_k(P,Q) \geq \int h\mathrm{d}[P-Q] > \int \tilde{f}\mathrm{d}[P-Q] - 2\|h-\tilde{f}\|_\infty > \frac{L}{4} > 0,$$

a contradiction! Hence $\int f\mathrm{d}P = \int f\mathrm{d}Q$ for all $f \in C(\mathcal{X})$.

Now we prove $P = Q$. For any measurable set $\mathcal{E} \subseteq \mathcal{X}$, with $\epsilon > 0$ fixed, we can find closed sets $\mathcal{K}_1, \mathcal{K}_2$ and open sets $\mathcal{F}_1, \mathcal{F}_2$ such that $\mathcal{K}_1 \subseteq \mathcal{E} \subseteq \mathcal{F}_1$, $P(\mathcal{F}_1 \backslash \mathcal{K}_1) < \epsilon$, $\mathcal{K}_2 \subseteq \mathcal{E} \subseteq \mathcal{F}_2$, $Q(\mathcal{F}_2\backslash\mathcal{K}_2) < \epsilon$. Let $\mathcal{K} = \mathcal{K}_1 \cup \mathcal{K}_2$ and $\mathcal{F} = \mathcal{F}_1 \cap \mathcal{F}_2$, we have $\mathcal{K} \subseteq \mathcal{E} \subseteq \mathcal{F}$ and

$$P(\mathcal{F}\backslash\mathcal{K}) < \epsilon,\ Q(\mathcal{F}\backslash\mathcal{K}) < \epsilon.$$

Using Urysohn's lemma, there exists continuous function $g \in C(\mathcal{X}, [0,1])$ such that $g(x) = 1$ for all $x \in \mathcal{K}$, and $g(x) = 0$ for all $x \in \mathcal{F}$. Then

$$\left|P(\mathcal{E}) - \int g\mathrm{d}P\right| \leq P(\mathcal{F}\backslash\mathcal{K}) < \epsilon,\ \left|Q(\mathcal{E}) - \int g\mathrm{d}Q\right| \leq Q(\mathcal{F}\backslash\mathcal{K}) < \epsilon.$$

Because $\int g\mathrm{d}P = \int g\mathrm{d}Q$, we have $|P(\mathcal{E}) - Q(\mathcal{E})| < 2\epsilon$. Since ϵ is arbitrarily chosen, $P(\mathcal{E}) = Q(\mathcal{E})$, which concludes the proof. $\square$

4.12 Proof of Lemma 3.5

Proof. We fix the measurable function $g \in \mathcal{G}$. By Fenchel's duality, we have

$$g(x)\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)} - f\left(\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)}\right) \leq f^*(g(x)).$$

Take integration with respect to P on both sides of the equation above, we have

$$\mathbb{E}_{Z\sim Q}[g(Z)] - D_f(Q\|P) \leq \mathbb{E}_{X\sim P}[f^*(g(X))].$$

Since g is arbitrarily chosen, we immediately conclude the inequality in Lemma 3.5. The supremum can be found when the derivative of (3.9) vanishes. $\square$

4.13 Proof of Theorem 3.9

Proof. For any measurable g on $\mathcal{X}$ with $\|g\|_\infty < 1$, the following inequalities hold uniformly on $\mathcal{X}$:

$$g \geq \log(1+g),\ -g \geq \log(1-g).$$

Then we have

$$\int g \mathrm{d}P - \int g \mathrm{d}Q \geq \int \log(1+g)\mathrm{d}P + \int \log(1-g)\mathrm{d}Q.$$

Use the variational representations of total variation and Jensen-Shannon divergence, we have

$$\begin{aligned}
d_{\mathrm{TV}}(P,Q) &= \frac{1}{2} \sup_{\|g\|_\infty \leq 1} \int g \mathrm{d}[P-Q] \\
&\geq \frac{1}{2} \sup_{\|g\|_\infty < 1} \int g \mathrm{d}[P-Q] \\
&\geq \frac{1}{2} \sup_{\|g\|_\infty \leq 1} \int \log(1+g)\mathrm{d}P - \int \log(1-g)\mathrm{d}Q = d_{\mathrm{JS}}(P,Q).
\end{aligned}$$

□

4.14 Proof of Theorem 3.10

Proof. Let $\mathcal{H}$ be the RKHS reproduced by k. For any $f \in \mathcal{H}$, Proposition 2.10 implies

$$\|f\|_\infty \leq \|k\|_\infty \|f\|_{\mathcal{H}}.$$

Plug in to the variational representation of total variation, we obtain

$$\begin{aligned}
d_{\mathrm{TV}}(P,Q) &= \frac{1}{2} \sup_{\|f\|_\infty \leq 1} \int f \mathrm{d}[P-Q] \\
&\geq \frac{1}{2} \sup_{f \in \mathcal{H}, \|f\|_\infty \leq 1} \int f \mathrm{d}[P-Q] \\
&\geq \frac{1}{2} \sup_{f \in \mathcal{H}, \|k\|_\infty \|f\|_{\mathcal{H}} \leq 1} \int f \mathrm{d}[P-Q] \\
&= \frac{1}{2\|k\|_\infty} \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} \int f \mathrm{d}[P-Q] = \frac{\gamma_k(P,Q)}{2\|k\|_\infty},
\end{aligned}$$

where the inequalities follow from the order of the sets in which the supremum is taken. Thus we conclude the proof. □

References

[1] Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* 41(5), 2263-2291.

[2] Mak, S. and Joseph, V. R. (2018). Support points. *Ann. Statist.* 46(6A), 2562-2592.

[3] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* 13 723–773.

[4] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines.* Springer, New York.

[5] Polyanskiy, Y. and Wu, Y. (2022). *Information Theory: From Coding to Learning (draft of October 20, 2022).* Cambridge University Press.