Locality-Sensitive Hashing for f-Divergences and Kreĭn Kernels: Mutual Information Loss and Beyond

Lin Chen[1,2], Hossein Esfandiari[1], Thomas Fu[1], and Vahab Mirrokni[1]

[1]Google Research, [2]Yale University

1. Locality-Sensitive Hashing (LSH)

$\mathcal{M}$: the universal set of items (the database), endowed with a distance function D.

Intuition of LSH: For any p and q in $\mathcal{M}$,

- If they are close, they are more likely to have the same hash value;
- If they are far apart, they are more likely to have different hash values.

(r_1, r_2, p_1, p_2)-sensitive LSH: Let $\mathcal{H} = \{h : \mathcal{M} \to U\}$ be a family of hash functions, where U is the set of possible hash values. Assume that there is a distribution $h \sim \mathcal{H}$ over the family of functions. This family $\mathcal{H}$ is called (r_1, r_2, p_1, p_2)-sensitive ($r_1 < r_2$ and $p_1 > p_2$) for D, if for $\forall p, q \in \mathcal{M}$ the following statements hold:

- If $D(p, q) \le r_1$, then $\Pr_{h\sim\mathcal{H}}[h(p) = h(q)] \ge p_1$;
- If $D(p, q) > r_2$, then $\Pr_{h\sim\mathcal{H}}[h(p) = h(q)] \le p_2$.

2. f-Divergence

- **f-divergence** from P to Q [5] is defined by is defined by
$$D_f(P \parallel Q) = \sum_{i\in\Omega} Q(i) f\left(\frac{P(i)}{Q(i)}\right) , \tag{1}$$
where $f : (0, \infty) \to \mathbb{R}$ be convex s.t. $f(1) = 0$. It is not symmetric in general: $D_f(P \parallel Q) \neq D_f(Q \parallel P)$.
- **KL-Divergence** $D_{\mathrm{KL}}(P \parallel Q)$ is the f_{KL}-divergence [4], where $f_{\mathrm{KL}}(t) = t \ln t + (1 - t)$. We have $D_{\mathrm{KL}}(P \parallel Q) = \sum_{i\in\Omega} P(i) \ln \frac{P(i)}{Q(i)}$.
- **Squared Hellinger Distance** $H^2(P, Q)$ is the hel-divergence [6], where $\mathrm{hel}(t) = \frac{1}{2}(\sqrt{t} - 1)^2$. We have $H^2(P, Q) = \frac{1}{2}\int_\Omega (\sqrt{dP} - \sqrt{dQ})^2$.
- **Triangular Discrimination:** If $\delta(t) = \frac{(t-1)^2}{t+1}$, the δ-divergence is the *triangular discrimination* (also known as Vincze-Le Cam distance) [9, 14]. If the sample space is finite, the triangular discrimination between P and Q is given by $\Delta(P \parallel Q) = \sum_{i\in\Omega} \frac{(P(i)-Q(i))^2}{P(i)+Q(i)}$.
- **Jensen-Shannon (JS) Divergence** is a symmetrized version of the KL divergence. If $P \ll Q$, $Q \ll P$ and $M = (P + Q)/2$, it is defined by
$$D_{\mathrm{JS}}(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M) . \tag{2}$$

3. Mutual Information Loss and Generalized JS Divergence

Suppose that two random variables X and C obeys a joint distribution $p(X, C)$. This joint distribution can model a dataset where X denotes the feature value of a data point and C denotes its label [2]. Let $\mathcal{X}$ and $\mathcal{C}$ denote the support of X and C (*i.e.*, the universal set of all possible feature values and labels), respectively. Consider clustering two feature values into a new combined value. This operation can be represented by the following map

$$\pi_{x,y} : \mathcal{X} \to \mathcal{X}\backslash\{x, y\} \cup \{z\} \quad \text{such that} \quad \pi_{x,y}(t) = \begin{cases} t, & t \in \mathcal{X} \setminus \{x, y\} \\ z, & t = x, y \end{cases} ,$$

where x and y are the two feature values to be clustered and $z \notin \mathcal{X}$ is the new combined feature value. To make the dataset after applying the map $\pi_{x,y}$ preserve as much information of the original dataset as possible, one has to select two feature values x and y such that the mutual information loss incurred by the clustering operation $\mathrm{mil}(x, y) = I(X; C) - I(\pi_{x,y}(X); C)$ is minimized, where $I(\cdot; \cdot)$ is the mutual information between two random variables [4]. Note that the *mutual information loss (MIL) divergence* $\mathrm{mil} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is symmetric in both arguments and always non-negative due to the data processing inequality [4]. Next, we motivate the generalized Jensen-Shannon divergence. If we let P and Q be the conditional distribution of C given $X = x$ and $X = y$, respectively, such that $P(c) = p(C = c|X = x)$ and $Q(c) = p(C = c|X = y)$, the mutual information loss can be re-written as

$$\lambda D_{\mathrm{KL}}(P \parallel M_\lambda) + (1 - \lambda) D_{\mathrm{KL}}(Q \parallel M_\lambda) , \tag{3}$$

where $\lambda = \frac{p(x)}{p(x)+p(y)}$ and the distribution $M_\lambda = \lambda P + (1 - \lambda)Q$. Note that (3) is a generalized version of (2). Therefore, we define the *generalized Jensen-Shannon (GJS) divergence* between P and Q [10, 1, 8] by $D^\lambda_{\mathrm{GJS}}(P \parallel Q) = \lambda D_{\mathrm{KL}}(P \parallel M_\lambda) + (1 - \lambda) D_{\mathrm{KL}}(Q \parallel M_\lambda)$, where $\lambda \in [0, 1]$ and $M_\lambda = \lambda P + (1 - \lambda)Q$. We immediately have $D^{1/2}_{\mathrm{GJS}}(P \parallel Q) = D_{\mathrm{JS}}(P \parallel Q)$, which indicates that the JS divergence is indeed a special case of the GJS divergence when $\lambda = 1/2$. The GJS divergence has another equivalent definition $D^\lambda_{\mathrm{GJS}}(P \parallel Q) = H(M_\lambda) - \lambda H(P) - (1 - \lambda)H(Q)$, where $H(\cdot)$ denotes the Shannon entropy [4]. In contrast to the MIL divergence, the GJS $D^\lambda_{\mathrm{GJS}}(\cdot \parallel \cdot)$ is not symmetric in general as the weight $\lambda \in [0, 1]$ is fixed and not necessarily equal to $1/2$.

4. Positive Definite Kernel and Kreĭn Kernel

Positive definite kernel [13] Let $\mathcal{X}$ be a non-empty set. A symmetric, real-valued map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel on $\mathcal{X}$ if for all positive integer n, real numbers $a_1, \ldots, a_n \in \mathbb{R}$, and $x, \ldots, x_n \in \mathcal{X}$, it holds that $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) \ge 0$.

A kernel is said to be a *Kreĭn* kernel if it can be represented as the difference of two positive definite kernels.

Kreĭn kernel [11] Let $\mathcal{X}$ be a non-empty set. A symmetric, real-valued map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Kreĭn kernel on $\mathcal{X}$ if there exists two positive definite kernels k_1 and k_2 on $\mathcal{X}$ such that $k(x, y) = k_1(x, y) - k_2(x, y)$ holds for all $x, y \in \mathcal{X}$.

5. LSH Schemes for f-Divergences

We build LSH schemes for f-divergences based on approximation via another f-divergence if the latter admits an LSH family. If D_f and D_g are two divergences associated with convex functions f and g as defined by (1), the approximation ratio of $D_f(P \parallel Q)$ to $D_g(P \parallel Q)$ is determined by the ratio of the functions f and g, as well as the ratio of P to Q (to be precise, $\inf_{i\in\Omega} \frac{P(i)}{Q(i)}$) [12].

Proposition 1[Proof in **??**] Let $\beta_0 \in (0, 1), L, U > 0$ and let f and g be two convex functions $(0, \infty) \to \mathbb{R}$ that obey $f(1) = 0$, $g(1) = 0$, and $f(t), g(t) > 0$ for every $t \neq 1$. Let $\mathcal{P}$ be a set of probability measures on a finite sample space Ω such that for every $i \in \Omega$ and $P, Q \in \mathcal{P}$, $0 < \beta_0 \le \frac{P(i)}{Q(i)} \le \beta_0^{-1}$. Assume that for every $\beta \in (\beta_0, 1) \cup (1, \beta_0^{-1})$, it holds that $0 < L \le \frac{f(\beta)}{g(\beta)} \le U < \infty$. If $\mathcal{H}$ forms an (r_1, r_2, p_1, p_2)-sensitive family for g-divergence on $\mathcal{P}$, then it is also an (Lr_1, Ur_2, p_1, p_2)-sensitive family for f-divergence on $\mathcal{P}$.

?? provides a general strategy of constructing LSH families for f-divergences. The performance of such LSH families depends on the tightness of the approximation. In **??** and **??**, as instances of the general strategy, we derive concrete results for the generalized Jensen-Shannon divergence and triangular discrimination, respectively.

Generalized Jensen-Shannon Divergence First, **??** shows that the GJS divergence is indeed an instance of f-divergence. lemma Define $m_\lambda(t) = \lambda t \ln t - (\lambda t + 1 - \lambda) \ln(\lambda t + 1 - \lambda)$. For any $\lambda \in [0, 1]$, $m_\lambda(t)$ is convex on $(0, \infty)$ and $m_\lambda(1) = 0$. Furthermore, m_λ-divergence yields the GJS divergence with parameter λ.

We choose to approximate it via the squared Hellinger distance, which plays a central role in the construction of the hash family with desired properties.

The approximation guarantee is established in theorem 1. We show that the ratio of $D^\lambda_{\mathrm{GJS}}(P \parallel Q)$ to $H^2(P, Q)$ is upper bounded by the function $U(\lambda)$ and lower bounded by the function $L(\lambda)$. Furthermore, theorem 1 shows that $U(\lambda) \le 1$, which implies that the squared Hellinger distance is an upper bound of the GJS divergence.

Theorem (Proof in ??)

We assume that the sample space Ω is finite. Let P and Q be two different distributions on Ω. For every $t > 0$ and $\lambda \in (0, 1)$, we have

$$L(\lambda) H^2(P, Q) \le D^\lambda_{\mathrm{GJS}}(P \parallel Q) \le U(\lambda) H^2(P, Q) \le H^2(P, Q),$$

where $L(\lambda) = 2 \min\{\eta(\lambda), \eta(1 - \lambda)\}$, $\eta(\lambda) = -\lambda \ln \lambda$ and $U(\lambda) = \frac{2\lambda(1-\lambda)}{1-2\lambda} \ln \frac{1-\lambda}{\lambda}$.

We show theorem 1 by showing a two-sided approximation result regarding m_λ and hel. This result might be of independent interest for other machine learning tasks, say, approximate information-theoretic clustering [3].

Lemma (Proof in ??)

Define $\kappa_\lambda(t) = \frac{m_\lambda(t)}{\mathrm{hel}(t)}$. For every $t > 0$ and $\lambda \in (0, 1)$, we have $\kappa_\lambda(t) = \kappa_{1-\lambda}(1/t)$ and $\kappa_\lambda(t) \in [L(\lambda), U(\lambda)]$.