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 set of items (the database), endowed with a distance 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 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 is defined by

$$D_f(P \parallel Q) = \sum_{i\in\Omega} Q(i) f\left(\frac{P(i)}{Q(i)}\right) , \tag{1}$$

for f convex and $f(1) = 0$. Generally $D_f(P \parallel Q) \neq D_f(Q \parallel P)$.

- **KL-Divergence**

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{i\in\Omega} P(i) \ln \frac{P(i)}{Q(i)}$$

is the f_{KL}-divergence, where $f_{\mathrm{KL}}(t) = t \ln t + (1 - t)$.

- **Squared Hellinger Distance (SHD)**

$$H^2(P,Q) = \frac{1}{2}\sum_{i\in\Omega}(\sqrt{P(i)} - \sqrt{Q(i)})^2$$

is the hel-divergence, where $\mathrm{hel}(t) = \frac{1}{2}(\sqrt{t} - 1)^2$.

- **Jensen-Shannon Divergence (JSD)** is a symmetrized version of the KL divergence. If $M = (P + Q)/2$, 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 (MIL)

Let $X \in \mathcal{X}$ be the feature value of a data item, $C \in \mathcal{C}$ be its label, and the joint distribution $p(X, C)$ model a dataset [1].

Consider clustering two feature values x, y into a new combined value z:

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

To make the dataset after clustering preserve as much information of the original dataset as possible, we want to minimize mutual information loss

$$\mathrm{mil}(x, y) = I(X; C) - I(\pi_{x,y}(X); C) ,$$

where $I(\cdot;\cdot)$ is the mutual information [3].

$\mathrm{mil}(x, y) = \mathrm{mil}(y, x) \ge 0$ due to the data processing inequality [3].

4. Generalized Jensen-Shannon Divergence (GJSD)

Let P and Q be the conditional distribution of C s.t.

$P(c) = p(C = c|X = x)$ and $Q(c) = p(C = c|X = y)$.

The MIL 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) with $\lambda = 1/2$. Therefore, for $\lambda \in [0, 1]$, we define the GJSD by

$$D^\lambda_{\mathrm{GJS}}(P \parallel Q) = \lambda D_{\mathrm{KL}}(P \parallel M_\lambda) + (1 - \lambda) D_{\mathrm{KL}}(Q \parallel M_\lambda) .$$

In contrast to the MIL divergence, the GJSD $D^\lambda_{\mathrm{GJS}}(\cdot \parallel \cdot)$ is not symmetric unless $\lambda = 1/2$.

Lemma 1 The GJSD is m_λ-divergence, where

$$m_\lambda(t) = \lambda t \ln t - (\lambda t + 1 - \lambda) \ln(\lambda t + 1 - \lambda) .$$

5. Positive Definite Kernel and Kreĭn Kernel

Positive definite (PD) kernel. A symmetric map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a PD kernel on $\mathcal{X}$ if for all $a_1, \dots, a_n \in \mathbb{R}$, and $x_1, \dots, x_n \in \mathcal{X}$, it holds that $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) \ge 0$.

Kreĭn kernel. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Kreĭn kernel on $\mathcal{X}$ if there exist PD kernels k_1 and k_2 s.t. $k(x, y) = k_1(x, y) - k_2(x, y)$.

6. 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.

Proposition 1 Let $\beta_0 \in (0, 1), L, U > 0$ and let f and g be convex and s.t. $f(1) = 0$, $g(1) = 0$, and $f(t), g(t) > 0$ for every $t \neq 1$. Let $\mathcal{P}$ be a set of probability measures on Ω s.t. 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 $\mathcal{P}$.

References

[1] MohammadHossein Bateni, Lin Chen, Hossein Esfandiari, Thomas Fu, Vahab S Mirrokni, and Afshin Rostamizadeh. Categorical feature compression via submodular optimization. *ICML*, 2019.

[2] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 380–388. ACM, 2002.

[3] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[4] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *SoCG*, pages 253–262. ACM, 2004.

[5] Behnam Neyshabur and Nathan Srebro. On symmetric and asymmetric lshs for inner product search. In *ICML*, pages 1926–1934, 2015.

[6] Anshumali Shrivastava and Ping Li. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). In *NeurIPS*, pages 2321–2329, 2014.

7. Example: LSH for Generalized Jensen-Shannon Divergence via Hellinger Approximation

We choose to approximate GJSD via the SHD.

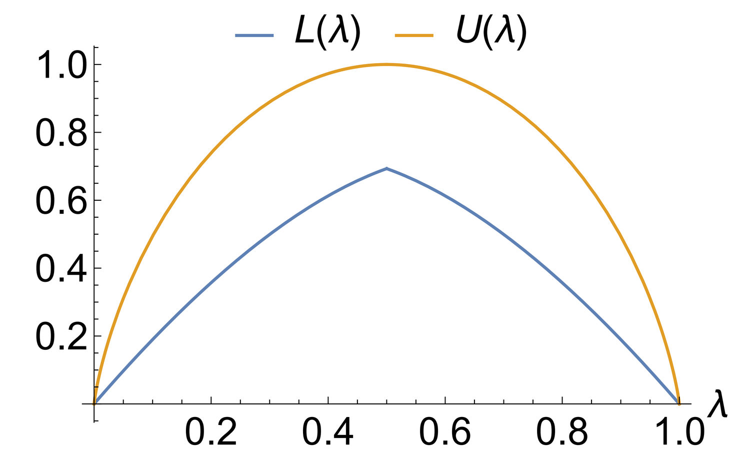
Theorem 1 For every $t > 0$ and $\lambda \in (0, 1)$, we have

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

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

Theorem 1 is based on the following two-sided approximation between m_λ and hel. This result might be of independent interest.

Lemma 2 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 $\kappa_\lambda(t) \in [L(\lambda), U(\lambda)]$.



We identify P and Q with vectors $[P(i)]_{i\in\Omega}, [Q(i)]_{i\in\Omega} \in \mathbb{R}^{|\Omega|}$. In this case, $H^2(P,Q) = \frac{1}{2}\|\sqrt{P} - \sqrt{Q}\|_2^2$, where $\sqrt{P} \triangleq [\sqrt{P(i)}]_{i\in\Omega}$ and $\sqrt{Q} \triangleq [\sqrt{Q(i)}]_{i\in\Omega}$. Therefore, the SHD can be endowed with the L^2-LSH family [4] applied to the square root of the vector. The LSH for the GJSD is

$$h_{\mathrm{a},b}(P) = \left\lceil \frac{\mathrm{a}^\top \sqrt{P} + b}{r} \right\rceil , \tag{4}$$

where $\mathrm{a} \sim \mathcal{N}(0, I)$, $b \sim \mathrm{Unif}[0, r]$, and $r > 0$.

Theorem 2 Let $c = \|\sqrt{P} - \sqrt{Q}\|_2$ and f_2 be the probability density function of the absolute value of $\mathcal{N}(0, 1)$. The hash functions $\{h_{\mathrm{a},b}\}$ defined in (4) form a $(R, c^2\frac{U(\lambda)}{L(\lambda)}R, p_1, p_2)$-sensitive family for the GJSD, where $R > 0$, $p_1 = p(1)$, $p_2 = p(c)$, and $p(u) = \int_0^r \frac{1}{u} f_2(t/u)(1 - t/r)dt$.

8. MIL is a Kreĭn Kernel

Recall that we assume a joint distribution $p(X, C)$. Let $x, y \in \mathcal{X}$ be represented by $\mathrm{x} = [p(c, x) : c \in \mathcal{C}] \in [0, 1]^{|\mathcal{C}|}$ and $\mathrm{y} = [p(c, y) : c \in \mathcal{C}] \in [0, 1]^{|\mathcal{C}|}$.

Theorem 4 $\mathrm{mil}(\mathrm{x}, \mathrm{y})$ is a Kreĭn kernel on $[0, 1]^{|\mathcal{C}|}$: $\mathrm{mil} = K_1 - K_2$, where $K_1(\mathrm{x}, \mathrm{y}) = k(\sum_{c\in\mathcal{C}} p(c, x), \sum_{c\in\mathcal{C}} p(c, y))$ and $K_2(\mathrm{x}, \mathrm{y}) = \sum_{c\in\mathcal{C}} k(p(c, x), p(c, y))$ are PD kernels, and $k(a, b) = a \ln \frac{a}{a+b} + b \ln \frac{b}{a+b}$.

To construct explicit feature maps for K_1 and K_2, we need **Lemma 3**.

Lemma 3 k is a PD kernel on $[0, 1]$ s.t. $k(x, y) = \langle \Phi(x), \Phi(y) \rangle \triangleq \int_{\mathbb{R}} \Phi_w(x)^* \Phi_w(y) dw$, where $\Phi_w(x) \triangleq e^{-iw\ln(x)} \sqrt{x\rho(w)}$, $\rho(w) = \frac{2\,\mathrm{sech}(\pi w)}{1+4w^2}$ and * denotes the complex conjugate.

9. Intuition of Kreĭn-LSH: Reduction to Maximum Inner Product Search (MIPS)

We reduce the problem of designing the LSH for a Kreĭn kernel to the problem of designing the LSH for MIPS [6].

Reduction to MIPS: If $K_i(x, y) = \langle \Psi_i(x), \Psi_i(y) \rangle$ $(i = 1, 2)$, then $K = K_1 - K_2$ can also represented as an inner product

$$K(x, y) = \langle \Phi_1(x) \oplus \Phi_2(x), \Phi_1(y) \oplus -\Phi_2(y) \rangle , \tag{5}$$

where $\oplus$ denotes the direct sum.

Apply LSH to MIPS As an example, we use Simple-LSH [5]. Assume $\|\mathrm{x}\|_2^2 \le M$ for all $\mathrm{x} \in \mathcal{M}$. For $\mathrm{x}, \mathrm{y} \in \mathcal{M}$, Simple-LSH performs the following transform

$$L_1(\mathrm{x}) \triangleq [\mathrm{x}, \sqrt{M - \|\mathrm{x}\|_2^2}, 0], L_2(\mathrm{y}) \triangleq [\mathrm{y}, 0, \sqrt{M - \|\mathrm{y}\|_2^2}] .$$

Since $\|L_1\|_2 = \|L_2\|_2 = M$, their cosine similarity is proportional to their inner product: $\frac{L_1(\mathrm{x})^\top L_2(\mathrm{y})}{\|L_1(\mathrm{x})\|\|L_2(\mathrm{y})\|} = \frac{L_1(\mathrm{x})^\top L_2(\mathrm{y})}{M^2}$. In fact, Simple-LSH is a reduction from MIPS to LSH for the cosine similarity. An LSH for the cosine similarity [2]

$$h(\mathrm{x}) \triangleq \mathrm{sign}(\mathrm{x}^\top L_i(\mathrm{x})), \quad \mathrm{a} \sim \mathcal{N}(0, I), i = 1, 2$$

can be used for MIPS and thereby LSH for the MIL via our reduction.

10. Algorithm: Kreĭn-LSH

To make the above intuition practical, we have to truncate and discretize the integral $k(x, y) = \int_R \Phi_w(x)^* \Phi_w(y) dw$.

Input: Discretization parameters $J \in \mathbb{N}$ and $\Delta > 0$.

1: $w_j \leftarrow (j - 1/2)\Delta$ for $j = 1, \dots, J$

2: Construct the atomic transform

$$\tau(x, w, j) \triangleq \left[\cos(w \ln(x))\sqrt{2x \int_{(j-1)\Delta}^{j\Delta} \rho(w')dw'}, \sin(w \ln(x))\sqrt{2x \int_{(j-1)\Delta}^{j\Delta} \rho(w')dw'}\right] .$$

3: Construct the left and right basic transform

$$\eta_1(\mathrm{x}) \triangleq \bigoplus_{j=1}^{J} \tau(p(x), w_j, j) \oplus \bigoplus_{j=1}^{J} \bigoplus_{c\in\mathcal{C}} \tau(p(c, x), w_j, j) ,$$

$$\eta_2(\mathrm{x}) \triangleq \bigoplus_{j=1}^{J} \tau(p(x), w_j, j) \oplus \bigoplus_{j=1}^{J} \bigoplus_{c\in\mathcal{C}} -\tau(p(c, x), w_j, j) .$$

4: Construct the left and right Kreĭn transform

$$T_1(\mathrm{x}, M) \triangleq [\eta_1, \sqrt{M - \|\eta_1(\mathrm{x})\|_2^2}, 0],$$

$$T_2(\mathrm{y}, M) \triangleq [\eta_2, 0, \sqrt{M - \|\eta_2(\mathrm{x})\|_2^2}] .$$

where M is a constant s.t. $M \ge \|\eta_1(\mathrm{x})\|_2^2 = \|\eta_2(\mathrm{x})\|_2$.

5: Sample $\mathrm{a} \sim \mathcal{N}(0, I)$ and construct the hash function $h(\mathrm{x}; M) \triangleq \mathrm{sign}(\mathrm{a}^\top T(\mathrm{x}, M))$, where T is either T_1 or T_2.