Lecture-20: Local behavior of divergence

1 Local behavior of divergence

KL divergence is in general not continuous. Nevertheless, it is reasonable to expect that in non-pathological cases the functional $D(P\|Q)$ vanishes when P approaches Q "smoothly". Due to the smoothness and strict convexity of $x \ln x$ at $x = 1$, it is then also natural to expect that this functional decays "quadratically". In this section, we examine this question first along the linear interpolation between P and Q, then, more generally, in smooth parametrized families of distributions. These properties will be extended to more general divergences later.

Definition 1.1. Consider a sample space Ω and event space $\mathcal{F}$. A probability measure $P : \mathcal{F} \to [0,1]$ satisfies σ-additivity and certainty axioms. For a random variable $X : \Omega \to \mathcal{X}$, we define set of measures for X as $\mathcal{M}(\mathcal{X})$ that consists of probability measures $P : \sigma(X) \to [0,1]$ that satisfies σ-additivity and certainty axioms.

1.1 Local behavior of divergence for mixtures

Consider a $\mathcal{F}$ measurable random variable $X : \Omega \to \mathcal{X}$ and two probability measures $P, Q \in \mathcal{M}(\mathcal{X})$. Let $\lambda \in [0,1], \bar{\lambda} \triangleq 1 - \lambda$ and consider $D(\lambda P + \bar{\lambda} Q \| Q)$, which vanishes as $\lambda \to 0$. Next, we show that this decay is always sublinear.

Lemma 1.2. *The map $h : \mathbb{R}_+ \to \mathbb{R}$ defined as $h(x) \triangleq x \ln x$ for all $x \in \mathbb{R}_+$, is convex.*

Proof. We observe that the second derivative of h exists and $h''(x) = \frac{1}{x} > 0$ for all $x \in \mathbb{R}_+$ □

Lemma 1.3. *We define a map $k_g : [0,1] \to \mathbb{R}$ as $k_g(\lambda) \triangleq (\lambda g + \bar{\lambda}) \ln(\lambda g + \bar{\lambda})$ for all $\lambda \in [0,1]$ and $g \in \mathbb{R}_+$. Then, the following statements hold true.*
(a) $k_g(0) = 0$ and $k_g(1) = g \ln g$ and $k_g(\lambda) \leqslant \lambda g \ln g$ for all $\lambda \in [0,1]$.
(b) k_g is a convex map in λ.
(c) $k_g(\lambda)/\lambda$ is increasing in $\lambda \in [0,1]$.

Proof. We define a Bernoulli random variable $X : \Omega \to \{g, 1\}$ with probability mass function $P_X(g) \triangleq P\{X = g\} = \lambda$, then we observe that $k_g(\lambda) = h(\mathbb{E}X)$.

(a) When $\lambda = 0$, we have $X = 1$ almost surely, resulting in $k_g(0) = h(\mathbb{E}X) = h(1) = 0$. When $\lambda = 1$, we have $X = f$ almost surely, resulting in $k_g(1) = h(\mathbb{E}X) = h(g) = g \ln g$. Applying Jensen inequality for convex map h, we get $k_g(\lambda) = h(\mathbb{E}X) \leqslant \mathbb{E}h(X) = \lambda g \ln g$ for all $\lambda \in [0,1]$.
(b) The result follows since $k_g''(\lambda) = \frac{(g-1)^2}{\lambda g + \bar{\lambda}} \geqslant 0$ for all $\lambda \in [0,1]$ and $f \in \mathbb{R}_+$. Alternatively, one can observe that $k_g(\lambda) = h(\mathbb{E}X)$ where $\mathbb{E}X = g\lambda + \bar{\lambda}$ is a composition of an affine and a convex map, and hence is convex.
(c) For the convex function k_g, we have $k_g(0) - k_g(\lambda) \geqslant -\lambda k_g'(\lambda)$. Rearranging, we get $(\lambda k_g'(\lambda) - k_g(\lambda))/\lambda^2 \geqslant k_g(0)/\lambda^2 = 0$. Recognizing that the left hand side of the previous equation is the first derivative of $k_g(\lambda)/\lambda$ with respect to λ, we get the result.

□

Definition 1.4. KL divergence between two binary distributions is denoted by $d(p\|q) \triangleq D((1-p, p)\|(1-q, q)) = (1-p)\log_2 \frac{1-p}{1-q} + p \log_2 \frac{p}{q}$ for all $p, q \in [0,1]$.

Definition 1.5 (Mixture distribution). For $\lambda \in [0,1]$ and $P, Q \in \mathcal{M}(\mathcal{X})$, we define a *mixture distribution* $P^\lambda \triangleq \lambda P + \bar{\lambda} Q \in \mathcal{M}(\mathcal{X})$.

Proposition 1.6. *For mixing parameter $\lambda \in [0,1]$ and $P, Q \in \mathcal{M}(\mathcal{X})$, the following are true for the first derivative of mixture distribution $P^\lambda \in \mathcal{M}(\mathcal{X})$ with respect to λ at 0.*

(a) If $D(P\|Q) < \infty$, then the one-sided derivative of $D(P^\lambda\|Q)$ at $\lambda = 0$ vanishes, i.e. $\frac{d}{d\lambda}\Big|_{\lambda=0} D(P^\lambda\|Q) = 0$.

(b) If we exchange the arguments, the criterion is even simpler,i.e. $P \ll Q$ iff $\frac{d}{d\lambda}\Big|_{\lambda=0} D(Q\|P^\lambda) = 0$.

Proof. Since $\lim_{\lambda\downarrow 0} D(P^\lambda\|Q) = 0$, we note that $\frac{d}{d\lambda}\Big|_{\lambda=0} D(P^\lambda\|Q) = \lim_{\lambda\downarrow 0}\frac{1}{\lambda}D(P^\lambda\|Q)$.

(a) Since $D(P\|Q) < \infty$, we have $P \ll Q$ and we define relative density $g \triangleq \frac{dP}{dQ}$. From the definition of KL divergence and definition of k_g in Lemma 1.3, we get

$$\frac{1}{\lambda}D(P^\lambda\|Q) = \mathbb{E}_Q\Big[\frac{1}{\lambda}(\lambda g + \bar{\lambda})\ln(\lambda g + \bar{\lambda})\Big] = \mathbb{E}_Q\frac{k_g(\lambda)}{\lambda}.$$

Recall that $k_g(\lambda)/\lambda \leqslant g\ln g$ is a monotone increasing and bounded map, where $\mathbb{E}_Q g\ln g = D(P\|Q) < \infty$. Thus, we can apply the monotone convergence theorem to interchange limits and expectation, to obtain

$$\lim_{\lambda\downarrow 0}\frac{1}{\lambda}D(\lambda P + \bar{\lambda}Q\|Q) = \mathbb{E}_Q\Big[\lim_{\lambda\downarrow 0}\frac{k_g(\lambda)}{\lambda}\Big] = \mathbb{E}_Q k_g'(0) = \mathbb{E}_Q(g-1) = 0.$$

(b) If $P \not\ll Q$, then there exists $E \in \sigma(X)$ such that $Q(E) = 0$ and $p \triangleq P(E) > 0$. Consider the binary output space $\mathcal{Y} \triangleq \{0,1\}$, and the processing $X \mapsto Y \triangleq \mathbb{1}_E(X)$ where $Y : \Omega \to \mathcal{Y}$. This processing leads to Markov kernel $P_{Y|X} : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ such that $P_{Y|X}(1 \mid x) = \mathbb{1}_{\{x\in E\}}$. For input distribution Q and mixture input distribution P^λ, the corresponding output distributions are

$$Q_Y \triangleq (1 - Q(E), Q(E)) = (1,0), \quad P_Y^\lambda \triangleq (1 - \lambda P(E) - \bar{\lambda}Q(E), \lambda P(E) + \bar{\lambda}Q(E)) = (1-\lambda p, \lambda p).$$

Applying data processing inequality for divergence to this processing kernel $P_{Y|X}$, we get

$$D(Q\|P^\lambda) \geqslant D(Q_Y\|P_Y^\lambda) = d(0\|\lambda p) = -\ln(1-\lambda p).$$

It follows that $\frac{d}{d\lambda}\Big|_{\lambda=0} D(Q\|P^\lambda) = p > 0$.

If $P \ll Q$, then we define relative density $g \triangleq \frac{dP}{dQ}$ and observe that $\ln\bar{\lambda} \leqslant \ln(\bar{\lambda} + \lambda g) \leqslant \lambda(g-1)$ from monotonicity of ln and the fact that $\ln(1+x) \leqslant x$ for each $x \in \mathbb{R}$. Dividing by λ and assuming $\lambda < \frac{1}{2}$ we get for some absolute constants $c_1 = 1, c_2 = 1 + \sup_{\lambda<0.5}\left|\frac{\ln\bar{\lambda}}{\lambda}\right|$

$$\left|\frac{1}{\lambda}\ln(\bar{\lambda} + \lambda g)\right| \leqslant |g-1| \vee \left|\frac{\ln\bar{\lambda}}{\lambda}\right| \leqslant g + 1 + \left|\frac{\ln\bar{\lambda}}{\lambda}\right| \leqslant c_1 g + c_2.$$

We recall that $\mathbb{E}_Q g = 1 < \infty$ It follows that $\left|\frac{1}{\lambda}\ln(\bar{\lambda} + \lambda g)\right|$ is Q integrable. Applying dominated convergence theorem to exchange limits and expectation, we get

$$\lim_{\lambda\downarrow 0}\frac{1}{\lambda}D(Q\|P^\lambda) = \lim_{\lambda\to 0} -\int_{\mathcal{X}} dQ\frac{1}{\lambda}\ln(\lambda g + \bar{\lambda}) = -\int_{\mathcal{X}} dQ\lim_{\lambda\downarrow 0}\frac{1}{\lambda}\ln(\lambda g + \bar{\lambda}) = \int dQ(1-g) = 0.$$

□

Remark 1. The main message of Proposition 1.6 is that the function $\lambda \mapsto D(P^\lambda\|Q)$ is $o(\lambda)$ as $\lambda \to 0$. In fact, in most cases it is quadratic in λ.

Exercise 1.7. Let $\lambda \in [0,1], i \in \{0,1\}$ and $P_i, Q_i \in \mathcal{M}(\mathcal{X})$, to define mixture distribution $Q_i^\lambda \triangleq \lambda Q_i + \bar{\lambda}P_i$. Show that under suitable technical conditions, the following equations hold

$$\frac{d}{d\lambda}\Big|_{\lambda=0} D(Q_0^\lambda\|P_1) = \mathbb{E}_{Q_0}\ln\frac{dP_0}{dP_1} - D(P_0\|P_1),$$
$$\frac{d}{d\lambda}\Big|_{\lambda=0} D(Q_1^\lambda\|Q_0^\lambda) = \mathbb{E}_{Q_1}\ln\frac{dP_1}{dP_0} - D(P_1\|P_0) + \mathbb{E}_{P_1}\Big[1 - \frac{dQ_0}{dP_0}\Big].$$

Lemma 1.8. *We observe that* $S_x \triangleq \int_0^1 \frac{s}{x(1-s)+s}ds = \frac{x\ln x - (x-1)}{(x-1)^2}$.

Proof. We observe that

$$S_x \triangleq \int_0^1 \frac{sds}{x(1-s)+s} = \frac{1}{x}\int_0^1 \frac{(x(s-1)-s+s)ds}{x(1-s)+s} + \frac{1}{1-x}\int_x^1 \frac{dy}{y} = -\frac{1}{x} + \frac{S_x}{x} + \frac{\ln x}{x-1}.$$

Rearranging the terms, we get the result. □

Proposition 1.9 (KL is locally χ^2 like). *For any $\lambda \in [0,1]$ and distribution $P, Q \in \mathcal{M}(\mathcal{X})$ we define mixture distribution $P^\lambda \triangleq \lambda P + \bar{\lambda} Q \in \mathcal{M}(\mathcal{X})$. Then,*

$$\liminf_{\lambda \downarrow 0} \frac{1}{\lambda^2} D(P^\lambda \| Q) = \frac{1}{2}\chi^2(P\|Q).$$

Proof. We recall that f divergence remains unchanged for a shift of type $a(x-1)$ for any $f:(0,\infty) \to \mathbb{R}_+$. Thus, we observe that for $f(x) \triangleq x\ln x - (x-1)$, we have

$$D_f(P\|Q) = \mathbb{E}_Q f\Big(\frac{dP}{dQ}\Big) = D(P\|Q).$$

(a) Applying Fatou's lemma, observing that $f(1) = 0$, using the L'Hospital rule to take limits, the fact that $f'(x) = \ln x$, and definition of χ^2 divergence, we obtain

$$\liminf_{\lambda \downarrow 0} \frac{1}{\lambda^2} D(P^\lambda\|Q) = \liminf_{\lambda \downarrow 0} \frac{1}{\lambda^2}\mathbb{E}_Q f(\bar{\lambda}+\lambda g) \geqslant \mathbb{E}_Q \liminf_{\lambda \downarrow 0}\frac{1}{\lambda^2} f(\bar{\lambda}+\lambda g) = \frac{f''(1)}{2}\mathbb{E}_Q(g-1)^2 = \frac{1}{2}\chi^2(P\|Q).$$

It follows that if $\chi^2(P\|Q) = \infty$ then so is $\frac{1}{\lambda^2}D(P^\lambda\|Q)$. Thus, we can assume that $\chi^2(P\|Q) < \infty$ without any loss of generality.

(b) We assume $\chi^2(P\|Q) < \infty$ and from the definition of S_x in Lemma 1.8, we observe that $S_x = \frac{f(x)}{(x-1)^2}$ and the integrand of S_x is positive and decreasing for $x \in (0,\infty)$. In particular, we have

$$0 \leqslant \frac{f(x)}{(x-1)^2} = \int_0^1 \frac{s}{x(1-s)+s} ds \leqslant \int_0^1 ds = 1.$$

Taking $x = \bar{\lambda} + \lambda g$ for $g \triangleq \frac{dP}{dQ}$ in the above inequality, we obtain $0 \leqslant \frac{1}{\lambda^2} f(\bar{\lambda}+\lambda g) \leqslant (g-1)^2$. Since $\mathbb{E}_Q(g-1)^2 = \chi^2(P\|Q) < \infty$, applying dominated convergence theorem to exchange limit and expectation, we obtain

$$\lim_{\lambda \downarrow 0}\frac{1}{\lambda^2}\mathbb{E}_Q f(\bar{\lambda}+\lambda g) = \mathbb{E}_Q \lim_{\lambda \downarrow 0}\frac{1}{\lambda^2} f(\bar{\lambda}+\lambda g) = \frac{f''(1)}{2}\mathbb{E}_Q(g-1)^2 = \frac{1}{2}\chi^2(P\|Q).$$

□