Lecture-22: f-divergences for parametrized families

1 f-divergences in parametric families: Fisher information

We have already previewed the fact that in parametric families of distributions, the Hessian of the KL divergence turns out to coincide with the Fisher information.

Example 1.1 (GLM). For one-dimensional GLM with unit variance, the observation distribution $P_t \triangleq \mathcal{N}(t,1)$ for $t \in \Theta = \mathcal{X} \triangleq \mathbb{R}$. Let $\mu \in \mathcal{M}(\mathcal{X})$ be the Lebesgue measure and for all $(x,t) \in \mathcal{X} \times \Theta$,

$$p_t(x) \triangleq \frac{dP_t}{d\mu}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2}.$$

From the definition of f-divergence and the fact that $\frac{dP_t}{dP_0}(X) = e^{-\frac{1}{2}(t^2-2tX)}$, we get

$$D_f(P_t\|P_0) = \mathbb{E}_{X\sim P_0} f\Big(\frac{dP_t}{dP_0}(X)\Big) = \mathbb{E}_{X\sim P_0} f\Big(e^{-\frac{1}{2}(t^2-2tX)}\Big).$$

(a) Recall that $D(P_t\|P_0) = \mathbb{E}_{X\sim P_t} \ln \frac{dP_t}{dP_0}(X)$, and hence $D(P_t\|P_0) = -\frac{1}{2}\mathbb{E}_{X\sim P_t}(t^2 - 2tX) = \frac{t^2}{2}$.

(b) Recall that $\mathrm{TV}(P_t,P_0) = 1 - \int_{\mathcal{X}} dx (p_t(x) \wedge p_0(x))$. For $t > 0$, we observe that $p_t(x) < p_0(x)$ iff $x < \frac{t}{2}$. In this case,

$$\lim_{t\to 0} \frac{1}{t}\mathrm{TV}(P_t,P_0) = \lim_{t\to 0} \frac{1}{t}(1 - \int_{x<\frac{t}{2}} dP_t(x) - \int_{x>\frac{t}{2}} dP_0(x)) = \lim_{t\to 0} \frac{1}{t} \int_{-\frac{t}{2}}^{\frac{t}{2}} dP_0(x) = \frac{1}{\sqrt{2\pi}}.$$

For $t < 0$, we observe that $p_t(x) < p_0(x)$ iff $x > \frac{t}{2}$, and $\lim_{t\to 0} \frac{1}{t}\mathrm{TV}(P_t,P_0) = -\frac{1}{\sqrt{2\pi}}$.

(c) Recall that $\chi^2(P_t\|P_0) = \int_{\mathcal{X}} d\mu(x) \frac{p_t^2(X)}{p_0(X)} - 1$, and hence

$$\chi^2(P_t\|P_0) = -1 + \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} dx e^{-(x-t)^2 + \frac{1}{2}x^2} = -1 + \frac{e^{t^2}}{\sqrt{2\pi}} \int_{\mathbb{R}} dx e^{-\frac{1}{2}(x-2t)^2} = e^{t^2} - 1.$$

Dividing by t^2 on both sides and taking limit $t \to 0$, we obtain $\lim_{t\to 0} \frac{1}{t^2}\chi^2(P_t\|P_0) = 1$.

(d) Recall that $H^2(P_t,P_0) = 2 - 2\int_{\mathcal{X}} dx \sqrt{p_t(x)p_0(x)}$, and hence

$$H^2(P_t,P_0) = 2 - \frac{2}{\sqrt{2\pi}} \int_{\mathcal{X}} dx e^{-\frac{1}{4}((x-t)^2+x^2)} = 2 - \frac{2e^{-\frac{t^2}{8}}}{\sqrt{2\pi}} \int_{\mathcal{X}} dx e^{-\frac{1}{2}(x-\frac{t}{2})^2} = 2(1 - e^{-\frac{t^2}{8}}).$$

Dividing by t^2 on both sides and taking limit $t \to 0$, we obtain $\lim_{t\to 0} \frac{1}{t^2}H^2(P_t,P_0) = \frac{1}{4}$.

(e) Recall that $\mathrm{LC}(P_t,P_0) = \frac{1}{2}\int_{\mathcal{X}} dx \frac{(p_t(x)-p_0(x))^2}{p_t(x)+p_0(x)} = 1 - 2\int_{\mathcal{X}} dx \frac{p_t(x)p_0(x)}{p_t(x)+p_0(x)}$, and hence

$$\mathrm{LC}(P_t,P_0) = 1 - \frac{2}{\sqrt{2\pi}} \int_{\mathcal{X}} dx e^{-\frac{x^2}{2}} \frac{e^{tx}}{e^{tx} + e^{\frac{t^2}{2}}} = \frac{1}{\sqrt{2\pi}} \int_{\mathcal{X}} dx e^{-\frac{x^2}{2}} \frac{e^{\frac{t^2}{2}} - e^{tx}}{e^{\frac{t^2}{2}} + e^{tx}}.$$

Dividing by t^2 on both sides and taking limit $t \to 0$, exchanging limits and integral using dominated convergence theorem, we obtain $\lim_{t\to 0} \frac{1}{t^2}\mathrm{LC}(P_t,P_0) = \frac{1}{4}$.

We can summarize these results as

$$D(P_t\|P_0) = \frac{t^2}{2} + o(t^2), \qquad \mathrm{TV}(P_t, P_0) = \frac{|t|}{\sqrt{2\pi}} + o(t), \qquad \chi^2(P_t\|P_0) = t^2 + o(t^2),$$
$$H^2(P_t, P_0) = \frac{t^2}{4} + o(t^2), \qquad \mathrm{LC}(P_t, P_0) = \frac{t^2}{4} + o(t^2), \qquad .$$

We can see that with the exception of TV, other f-divergences behave quadratically under small displacement $t \to 0$. This turns out to be a general fact, and furthermore the coefficient in front of t^2 is given by the Fisher information at $t = 0$.

Definition 1.2 (Regular single-parameter families). Fix $\tau > 0$, parameter space $\Theta \triangleq [0, \tau)$, observation space $\mathcal{X}$, and a stochastic model $\mathcal{P}(\Theta) \triangleq \{P_t \in \mathcal{M}(\mathcal{X}) : t \in \Theta\}$. We define the following types of conditions that we call regularity at $t = 0$.

(a) $P_t \ll \mu$ for a fixed $\mu \in \mathcal{M}(\mathcal{X})$, and hence there exists relative density $p_t \triangleq \frac{dP_t}{d\mu}$ which is a measurable map $(t, x) \mapsto p_t(x) \in \mathbb{R}_+$.

(b) There exists a measurable function $(s, x) \mapsto \dot{p}_s(x)$ for all $(s, x) \in \Theta \times \mathcal{X}$, such that for μ-almost every x we have $\int_\Theta |\dot{p}_s(x)|\, ds < \infty$ and $p_t(x) = p_0(x) + \int_0^t \dot{p}_s(x) ds$ for every $t \in \Theta$. Furthermore, for μ-almost every x we have $\lim_{t \downarrow 0} \dot{p}_t(x) = \dot{p}_0(x)$.

(c) We have $\dot{p}_t(x) = 0$ whenever $p_0(x) = 0$ and, furthermore, $\int_\mathcal{X} d\mu(x) \sup_{t \in \Theta} \frac{(\dot{p}_t(x))^2}{p_0(x)} < \infty$.

(d) There exists a measurable function $(s, x) \mapsto \dot{h}_s(x)$ for all $(s, x) \in \Theta \times \mathcal{X}$, such that for μ-almost every x we have $\int_\Theta |\dot{h}_s(x)|\, ds < \infty$ and $h_t(x) \triangleq \sqrt{p_t(x)} = \sqrt{p_0(x)} + \int_0^t \dot{h}_s(x) ds$ for every $t \in \Theta$ Furthermore, for μ-almost every x we have $\lim_{t \downarrow 0} \dot{h}_t(x) = \dot{h}_0(x)$.

(e) The family of functions $\{(\dot{h}_t(x))^2 : t \in \Theta\}$ is uniformly μ-integrable.

Remark 1. Recall that the uniform integrability condition (e) is implied by the following stronger but easily verifiable condition

$$\int_\mathcal{X} d\mu(x) \sup_{t \in \Theta} (\dot{h}_t(x))^2 < \infty.$$

Recall that $\dot{h}_t(x) = \frac{d}{dt}\sqrt{p_t(x)}$ and Fisher information $J_F(t) = 4\int_\mathcal{X} d\mu(x)(\dot{h}_t(x))^2$ under suitable regularity conditions. If one also assumes the continuous differentiability of h_t then the uniform integrability condition becomes equivalent to the continuity of the Fisher information $t \mapsto J_F(t)$.

Lemma 1.3 (Taylor's integral remainder). *Let $f : \mathbb{R} \to \mathbb{R}$ be a function that has $k + 1$ continuous derivatives in some neighborhood U of $x = a$, then for any $x \in U$*

$$f(x) = \sum_{i=0}^{k} f^{(i)}(a) \frac{(x-a)^i}{i!} + \int_a^x f^{(k+1)}(t) \frac{(x-t)^k}{k!} dt.$$

Proof. We will show this by induction on k. The base case of $k = 0$ holds true from the fundamental theorem of calculus, that implies

$$f(x) = f(a) + \int_a^x f'(t) dt.$$

We assume that the hypothesis is true for k, and using the hypothesis and integration by parts

$$\begin{aligned} f(x) &= \sum_{i=0}^{k} f^{(i)}(a) \frac{(x-a)^i}{i!} - f^{(k+1)}(t) \frac{(x-t)^{k+1}}{(k+1)!}\Big|_{t=a}^{t=x} + \int_a^x f^{(k+2)}(t) \frac{(x-t)^{k+1}}{(k+1)!} dt \\ &= \sum_{i=0}^{k+1} f^{(i)}(a) \frac{(x-a)^i}{i!} + \int_a^x f^{(k+2)}(t) \frac{(x-t)^{k+1}}{(k+1)!} dt. \end{aligned}$$

Thus, we have shown that the inductive step holds true. □

Theorem 1.4. *Consider $\tau > 0$, the parameter space $\Theta \triangleq [0, \tau)$, and observation space $\mathcal{X}$. If the family $\mathcal{P}(\Theta) \triangleq \{P_t \in \mathcal{M}(\mathcal{X}) : t \in \Theta\}$ satisfies the conditions* (a), (b), *and* (c) *in Definition 1.2, then*

$$J_F(t) \triangleq \int_\mathcal{X} d\mu(x) \frac{(\dot{p}_t(x))^2}{p_t(x)} < \infty, \qquad \chi^2(P_t\|P_0) = J_F(0)t^2 + o(t^2), \qquad D(P_t\|P_0) = \frac{1}{2} J_F(0) t^2 + o(t^2).$$

Proof. We will show this in three parts.

(a) We can write the score $V \triangleq \nabla_t \ln p_t(X) = \frac{\dot{p}_t(X)}{p_t(X)}$, and hence the Fisher information matrix is

$$J_F(t) \triangleq \mathbb{E}_{X\sim P_t} VV^T = \int_{\mathcal{X}} d\mu(x) \frac{\dot{p}_t(x)^2}{p_t(x)}.$$

Condition (c) implies that $\sup_{t\in\Theta} \frac{\dot{p}_t(x)^2}{p_t(x)}$ is μ integrable, and hence $J_F(t) < \infty$ for all parameters $t \in \Theta$. Further, we can apply dominated convergence theorem to exchange limit and integral to obtain $J_F(0) = \int_{\mathcal{X}} d\mu(x) \frac{\dot{p}_0(x)^2}{p_0(x)}$.

(b) From Condition (c), we see that for any $x \in \{z \in \mathcal{X} : p_0(z) = 0\}$ we must have $\dot{p}_t(x) = 0$ and thus $p_t(x) = 0$ for all $t \in \Theta$. Hence, we may restrict all integrals below to P_0 almost sure subset $\{x \in \mathcal{X} : p_0(x) > 0\}$, on which the ratio $\frac{(p_t(x)-p_0(x))^2}{p_0(x)}$ is well-defined. Consequently, we have

$$\frac{1}{t^2}\chi^2(P_t\|P_0) = \int_{\mathcal{X}} d\mu(x) \frac{1}{p_0(x)} \Big(\frac{p_t(x) - p_0(x)}{t}\Big)^2. \tag{1}$$

By the continuity assumption in (b) we have $\lim_{t\to 0} \frac{p_t(x)-p_0(x)}{t} = \dot{p}_0(x)$ for every x. Furthermore, we also have $\Big|\frac{\dot{p}_u(x)\dot{p}_v(x)}{p_0(x)}\Big| \leqslant \sup_{t\in\Theta} \frac{(\dot{p}_t(x))^2}{p_0(x)}$, which is μ-integrable by (c). This implies that

$$\int_{\mathcal{X}} d\mu(x) \frac{\dot{p}_u(x)\dot{p}_v(x)}{p_0(x)} = \mathbb{E}_{X\sim P_0} \frac{\dot{p}_u(X)\dot{p}_v(X)}{p_0(X)^2} < \infty.$$

For $t \in \Theta$, we can apply Fubini theorem to exchange integrals, and write

$$\int_0^t du \int_0^t dv \mathbb{E}_{X\sim P_0} \frac{\dot{p}_u(X)\dot{p}_v(X)}{p_0(X)^2} = \mathbb{E}_{X\sim P_0}\Big(\int_0^t du \frac{\dot{p}_u(X)}{p_0(X)}\Big)^2 = \mathbb{E}_{X\sim P_0}\Big(\frac{p_t(X) - p_0(X)}{p_0(X)}\Big)^2 < \infty.$$

Applying the dominated convergence theorem to exchange limit and the integral in (1), we obtain

$$\lim_{t\to 0} \frac{1}{t^2}\chi^2(P_t\|P_0) = \int_{\mathcal{X}} d\mu(x) \frac{\dot{p}_0(x)^2}{p_0(x)} = J_F(0).$$

(c) We next show that for any f-divergence with twice continuously differentiable f without assuming (c),

$$\liminf_{t\downarrow 0} \frac{1}{t^2} D_f(P_t\|P_0) \geqslant \frac{f''(1)}{2} J_F(0). \tag{2}$$

Without any loss of generality, we can assume $f'(1) = f(1) = 0$ for the f in definition of f-divergence. This is due to the fact that $D_g = D_f$ for $g(x) \triangleq f(x) - c(x-1)$ and $g'(1) = 0$ for $c = f'(1)$. Applying Taylor integral remainder (Lemma 1.3) to function f for $k = 1$ terms in the neighborhood of $a = 1$, evaluated at $x = 1 + u$, and substituting $z \triangleq \frac{(t-1)}{u}$, we get

$$f(1+u) = f(1) + uf'(1) + u^2 \int_0^1 (1-z) f''(1+uz) dz.$$

Taking $u \triangleq \frac{(p_t(X)-p_0(X))}{p_0(X)}$, we recognize that $(1+u) = \frac{dP_t}{dP_0}(X)$. Substituting this u in the Taylor integral remainder of $f(1+u)$ in the neighborhood of 1, we can write

$$\frac{1}{t^2} D_f(P_t\|P_0) = \frac{1}{t^2}\mathbb{E}_{X\sim P_0} f\Big(\frac{dP_t}{dP_0}(X)\Big) = \int_0^1 dz(1-z)\mathbb{E}_{X\sim P_0}\left[f''\Big(1+z\frac{p_t(X)-p_0(X)}{p_0(X)}\Big)\Big(\frac{p_t(X)-p_0(X)}{tp_0(X)}\Big)^2\right].$$

From (b), we observe that $\lim_{t\downarrow 0} \frac{p_t(X)-p_0(X)}{tp_0(X)} = \frac{\dot{p}_0(X)}{p_0(X)}$ almost surely, and thus we have the following almost sure limit from the smoothness of f,

$$\lim_{t\downarrow 0} f''\Big(1+z\frac{p_t(X)-p_0(X)}{p_0(X)}\Big)\Big(\frac{p_t(X)-p_0(X)}{tp_0(X)}\Big)^2 = f''(1)\Big(\frac{\dot{p}_0(X)}{p_0(X)}\Big)^2.$$

Applying Fatou's lemma and the above limit to the expression for $\frac{1}{t^2}D_f(P_t\|P_0)$, we obtain (2). If $(1-z)f''\Big(1+z\frac{p_t(X)-p_0(X)}{p_0(X)}\Big)\Big(\frac{p_t(X)-p_0(X)}{tp_0(X)}\Big)^2$ is P_0-intgerable, then we can apply dominated convergence theorem to obtain

$$\liminf_{t\downarrow 0}\frac{1}{t^2}D_f(P_t\|P_0)=\frac{f''(1)}{2}J_F(0).$$

For $f(x)\triangleq x\ln x$ we have $f''(x)=\frac{1}{x}$, and hence

$$(1-z)f''\Big(1+z\frac{p_t(X)-p_0(X)}{p_0(X)}\Big)\Big(\frac{p_t(X)-p_0(X)}{tp_0(X)}\Big)^2\leqslant\sup_{t\in\Theta}\Big(\frac{\dot{p}_t(X)}{p_0(X)}\Big)^2.$$

Condition (c) implies that right hand side is P_0 integrable and hence the result follows.

□

Remark 2. Theorem 1.4 can be extended to the case of multi-dimensional parameters in the following fashion. Recall that the Fisher information matrix at parameter $\theta\in\mathbb{R}^d$ can be defined as

$$J_F(\theta)\triangleq\int_{\mathcal{X}}d\mu(x)(\nabla_\theta\sqrt{p_\theta(x)})(\nabla_\theta\sqrt{p_\theta(x)})^T.$$

We can derive $\chi^2(P_\theta\|P_0)=\theta^TJ_F(0)\theta+o(\|\theta\|^2)$ and $D(P_\theta\|P_0)=\frac{1}{2}\theta^TJ_F(0)\theta+o(\|\theta\|^2)$.

Remark 3. Theorem 1.4 applies to many cases, e.g. to smooth subfamilies of exponential families, for which one can take $\mu=P_0$ and $p_0(x)\triangleq 1$, but it is not sufficiently general.

Example 1.5 (Location families with compact support). We say that family P_t is a scalar location family if $\mathcal{X}=\mathbb{R}$, μ is the Lebesgue measure and $p_t(x)=p_0(x-t)$. For $\alpha>-1$, consider the following definition of $p_0(x)$ for all $x\in\mathcal{X}$,

$$p_0(x)\triangleq C_\alpha x^\alpha\mathbb{1}_{[0,1]}(x)+C_\alpha(2-x)^\alpha\mathbb{1}_{[1,2]}(x),$$

with C_α chosen appropriately from normalization. In this case, $\frac{(\dot{p}_t(x))^2}{p_0(x)}=\sup_{t\in[0,2)}\frac{(p_0'(x-t))^2}{p_0(x)}$, and we observe that condition (c) is not satisfied. Further both $\chi^2(P_t\|P_0)$ and $D(P_t\|P_0)$ are infinite for $t>0$, since $P_t\not\ll P_0$. But $J_F(0)<\infty$ whenever $\alpha>1$ and thus one expects that a certain remedy should be possible. Indeed, one can compute those f-divergences that are finite for $P_t\not\ll P_0$ and find that for $\alpha>1$ they are quadratic in t. In particular,

$$H^2(P_t,P_0)=\Theta(t^{1+\alpha})\mathbb{1}_{[0,1)}(\alpha)+\Theta(t^2\ln\frac{1}{t})\mathbb{1}_{\{\alpha=1\}}+\Theta(t^2)\mathbb{1}_{(1,\infty)}(\alpha).$$

$H^2(P_t,P_0)$ can be computed directly, or from a more general results of [221, Theorem VI.1.1][1]. For a relation between Hellinger and Fisher information see also (VI.5).

Remark 4. The previous example suggests that quadratic behavior as $t\downarrow 0$ can hold even when $P_t\not\ll P_0$, which is the case handled by the next (more technical) result. One can verify that condition (e) is indeed satisfied for all $\alpha>1$ in the previous example, thus establishing the quadratic behavior. Also note that the stronger uniform integrability condition only applies to $\alpha>2$.

Theorem 1.6. *Given a family of distributions $\{P_t:t\in[0,\tau)\}$ satisfying the conditions (a), (d), and* (e) *of Definition 1.2, we have*

$$\chi^2(P_t\|P_0)=t^2\bar{\epsilon}^2\Big(J_F(0)+\frac{1-4\epsilon}{\epsilon}J^{\#}(0)\Big)+o(t^2),\ \textit{for all}\ \epsilon\in(0,1)\qquad H^2(P_t,P_0)=\frac{t^2}{4}J_F(0)+o(t^2),$$

where $J_F(0)=4\int\dot{h}_0^2d\mu<\infty$ is the Fisher information and $J^{\#}(0)=\int\dot{h}_0^2\mathbb{1}_{\{h_0=0\}}d\mu$ is called the Fisher defect *at $t=0$.*

[1]Statistical significance of this calculation is that if we were to estimate the location parameter t from n *i.i.d.* observations, then precision δ_n^* of the optimal estimator up to constant factors is given by solving $H^2(P_{\delta_n^*},P_0)\asymp\frac{1}{n}$, cf. [221, Chapter VI]. For $\alpha<1$ we have $\delta_n^*\asymp n^{-1/(1+\alpha)}$ which is notably better than the empirical mean estimator (attaining precision of only $n^{-1/2}$). For $\alpha=1/2$ this fact was noted by D. Bernoulli in 1777 as a consequence of his (newly proposed) maximum likelihood estimation.