Lecture 1

Yury Polyanskiy

January 7, 2020
Typed by Suzanne Sigalla (ENSAE, CREST)

This first lecture will be about f-divergences and their applications in classical statistics. We introduce different definitions for f-divergences, from the most restrictive to the most general.

Definition 1 : let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function such that $f(1) = 0$. For two p.m.f. P, Q, we define the f-divergence between P and Q by:

$$D_f(P\|Q) = \sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right)$$

Definition 2 : in the case where $P << Q$ i.e. $\forall E, Q(E) = 0 \to P(E) = 0$, we may define the f-divergence between P and Q by:

$$D_f(P\|Q) = \int_x \mathrm{d}Q f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)$$

where we denote by $\frac{\mathrm{d}P}{\mathrm{d}Q}$ the Radon-Nikodym derivative of P relative to Q.

Definition 3 : let μ be any positive measure on $\mathscr{X}$ and suppose $\mathrm{d}P = p(x)\,\mathrm{d}\mu$, $\mathrm{d}Q = q(x)\,\mathrm{d}\mu$. Then, we may define the f-divergence between P and Q by:

$$D_f(P\|Q) = \int_{\{q>0\}} \mathrm{d}\mu\ q(x) f\left(\frac{p(x)}{q(x)}\right) + f'(\infty) P\left[q = 0\right]$$

***Remark* 1 -** Gelfand-Yaglom-Perez theorem ([GI59], [Per59]) states that:

$$\begin{aligned} D_f(P\|Q) &= \sup_{\varepsilon} D_f(P_{|\varepsilon}\|Q_{|\varepsilon}) \\ &= \sup_{\pi} \sum_{k=1}^{m} P\left(E_k\right) \log \frac{P\left(E_k\right)}{Q\left(E_k\right)} \end{aligned}$$

where the supremum is taken over all finite measurable partitions $\pi = \{E_1, \ldots, E_m\}$ $(m \geq 1)$ of $\mathscr{X}$.

In this lecture, we will work with the Definition 1.

Examples:

– The total variation distance, denoted by $TV(P, Q)$ is a f-divergence with:

$$f(x) = \frac{1}{2}|x - 1|$$

As pointed out by its name, the total variation distance is a distance.

– The Kullback-Leibler divergence, denoted by $D(P\|Q)$, is a f-divergence with:

$$f(x) = x \log x$$

The Kullback-Leibler divergence is not a distance ; it does not satisfy the symmetry condition.

- The chi-square divergence, denoted by $\chi^2(P\|Q)$, is a f-divergence with:

$$f(x) = (x-1)^2$$

We also remind that $\chi^2(P\|Q)$ may be written as:

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

The chi-square divergence is not a distance ; it does not satisfy the symmetry condition.

- The Hellinger-squared divergence, denoted by $H^2(P,Q)$,is a f-divergence with:

$$f(x) = (\sqrt{x}-1)^2$$

We remind that $H^2(P,Q)$ may be written as:

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

The Hellinger-squared divergence can be written as the square of a distance.

- The Symmetric Kullback-Leibler divergence, defined by $D_{SKL}(P\|Q) = D(P\|Q) + D(Q\|P)$, is a f-divergence with:

$$f(x) = x\log x - \log x$$

Note that even if D_{SKL} is symmetric, it still is not a distance.

- We have that:
  - $\sqrt{\chi^2(P\|\frac{P+Q}{2}) + \chi^2(Q\|\frac{P+Q}{2})}$;
  - $\sqrt{D(P\|\frac{P+Q}{2}) + D(Q\|\frac{P+Q}{2})}$

  both define a distance.

Theorem 1 (Main inequality)**.** *With the same hypothesis on f, P, Q as in Definition 1, we have:*

$$D_f(P\|Q) \geq 0$$

Proof.

$$D_f(P\|Q) = \sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right)$$

$$\overset{\text{Jensen}}{\geq} f\left(\sum_x \frac{P(x)Q(x)}{Q(x)}\right)$$
$$= f(1) = 0$$

□

***Remark* 2 -** WLOG, we may suppose $f'(1) = 0$.

Theorem 2 (Monotonicity)**.** *Denoting by A, B two real random variables and with the same hypothesis on f, P, Q as in Definition 1, we have:*

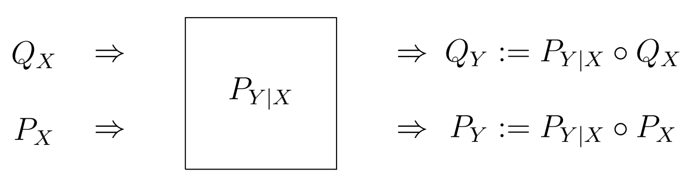
$$D_f(P_{A,B} \| Q_{A,B}) \geq D_f(P_A \| Q_A)$$

Proof.

$$\begin{aligned} D_f(P_{A,B} \| Q_{A,B}) &= \sum_{a,b} Q_{A,B}(a,b) f\left(\frac{P_{A,B}(a,b)}{Q_{A,B}(a,b)}\right) \\ &= \sum_a Q_A(a) \sum_b Q_{B|A}(b|a) f\left(\frac{P_{B|A}(b|a)P_A(a)}{Q_{B|A}(b|a)Q_A(a)}\right) \\ &\overset{\text{Jensen}}{\geq} \sum_a Q_A(a) f\left(\frac{P_A(a)}{Q_A(a)}\right) \end{aligned}$$

□

This drawing gives intuition about the following theorem:



Theorem 3 (Data Processing Inequality, DPI)**.** *Denoting by X, Y two real random variables and with the same hypothesis on f, P, Q as in Definition 1, we have:*

$$D_f(P_X \| Q_X) \geq D_f(P_Y \| Q_Y)$$

Proof.

$$D_f(P_{X,Y} \| Q_{X,Y}) = \sum_{x,y} Q_{X,Y}(x,y) f\left(\frac{P_{X,Y}(x,y)}{Q_{X,Y}(x,y)}\right)$$

Since :

$$\frac{P_{X,Y}(x,y)}{Q_{X,Y}(x,y)} = \frac{P_X(x)P_{Y|X}(y|x)}{Q_X(x)P_{Y|X}(y|x)} = \frac{P_X(x)}{Q_X(x)}$$

Therefore, this last ratio does not depend on y. It leads to:

$$\begin{aligned} D_f(P_{X,Y}\|Q_{X,Y}) &= \sum_x Q_X(x) f\left(\frac{P_X(x)}{Q_X(x)}\right) \\ &= D_f(P_X\|Q_X) \end{aligned}$$

Using that $D_f(P_{X,Y}\|Q_{X,Y}) \geq D_f(P_Y\|Q_Y)$ concludes the proof. □

Simple applications:

We fix P, Q as stated in Definition 1, A a subset of $\mathscr{X}$ and we define $Y(\omega) = \mathbb{1}_A(\omega)$.

1. $|P(A)-Q(A)| \leq TV(P,Q)$. Indeed, $|P(A)-Q(A)|$ can be seen as $TV\left[\text{Ber}(P(A)), \text{Ber}(Q(A))\right]$, where $\text{Ber}(p)$ designates a Bernouilli of parameter p.
2. $|P(A) - Q(A)| \leq \sqrt{\chi^2(P\|Q)Q(A)}$;
3. $|\sqrt{P(A)} - \sqrt{Q(A)}| \leq \sqrt{H^2(P,Q)}$;
4. $P(A)\log\frac{1}{Q(A)} \leq D(P\|Q) + \log 2$. This last point may give results of the following form, where $(P_n), (Q_n)$ denote sequences of distributions satisfying the usual assumptions, and (A_n) denotes a sequence of subsets of $\mathscr{X}$, such that $P_n(A_n) \to 1$.

$$Q_n(A_n) \geq \frac{1}{2}\exp\left[-D(P_n\|Q_n)(1+o(1))\right]$$

Theorem 4 (Convexity of D_f)**.** *With the same hypothesis on f as in Definition 1, the application $(P,Q) \mapsto D_f(P\|Q)$ is convex.*

Proof. let $\lambda \in (0,1)$ and $B \sim \text{Ber}(\lambda)$. We denote by $P_{X|B=0} = P_0, P_{X|B=1} = P_1, Q_{X|B=0} = Q_0, Q_{X|B=1} = Q_1$. We have $\mathbb{P}(B=0) = 1-\lambda := \overline{\lambda}$ and $\mathbb{P}(B=1) = \lambda$. We have:

$$\begin{aligned} D_f(P_{X,B}\|Q_{X,B}) &= \sum_{x,b} Q_{X,B}(x,b) f\left(\frac{P_{X,B}}{Q_{X,B}}\right) \\ &= \lambda D_f(P_1\|Q_1) + \overline{\lambda} D_f(P_0\|Q_0) \\ &\overset{\text{monotonicity/DPI}}{\geq} D_f(P_X\|Q_X) = D_f(\lambda P_1 + \overline{\lambda} P_0 \| \lambda Q_1 + \overline{\lambda} Q_0) \end{aligned}$$

which concludes the proof. □

***Remark* 3 -** Monotonicity is equivalent to DPI, which therefore implies convexity.

Corollary 1. *We fix Q. Then, with the same hypothesis as in Definition 1, the application $P \mapsto D_f(P\|Q)$ is convex.*

We would like to introduce an analog of functions' convex conjugate for distributions. We remind of the definition of convex conjugate for functions:

$$f_{\text{ext}}^*(y) = \sup_{x\in\mathbb{R}} \left[xy - f_{\text{ext}}(x)\right]$$

where f_{ext} is a convex extension of a convex function f to all $\mathbb{R}$. It is possible to consider:

$$\psi^*(g) = \sup_P \mathbb{E}_\rho(g) - D_{f_{\text{ext}}}(P\|Q)$$

where the supremum is taken over all signed measures.

$$\psi^*(g) = \sup_P \sum_x P(x)g(x) - Q(x) f_{\text{ext}}\left(\frac{P(x)}{Q(x)}\right)$$

Re-parametrizing $P(x) = y(x)Q(x)$:

$$\begin{aligned}\psi^*(g) &= \sup_{y(x)} \sum_x Q(x)\left[y(x)g(x) - f_{\text{ext}}\left[y(x)\right]\right] \\ &= \sum_x Q(x) \sup_y \left[yg(x) - f_{\text{ext}}(y)\right] \\ &= \mathbb{E}_Q f_{\text{ext}}^* \left[g(X)\right]\end{aligned}$$

Theorem 5. *With the same hypothesis as in Definition 1, the following holds for any f_{ext} such that $f_{ext} = f(x)$ for all $x > 0$:*

$$D_f(P\|Q) = \sup_g \left\{\mathbb{E}_P\left[g(x)\right] - \mathbb{E}_Q\left[f_{ext}^*\left[g(x)\right]\right]\right\}$$

where the supremum is taken over the set $\{g : \mathbb{R} \mapsto dom(f_{ext}^*)\}$.

Observation: e.g. $f_{\text{ext}} = \begin{cases} f(x) & x > 0 \\ +\infty & x \leq 0 \end{cases}$

Proof. "Almost rigorous proof":

$$\begin{aligned}D_f(P\|Q) &= \sum_x Q(x) \sup_g g\frac{P(x)}{Q(x)} - f_{\text{ext}}^*(g) \\ &= \sup_{g(x)} \sum_x g(x)P(x) - f_{\text{ext}}^*\left[g(x)\right]Q(x)\end{aligned}$$

□

Examples:

1. Kullback-Leibler:

$$f_{\text{ext}}(x) = \begin{cases} x \log x & x > 0 \\ +\infty & x \leq 0 \end{cases}$$
$$f^*_{\text{ext}}(y) = e^{y-1}$$

Then:

$$\begin{aligned} D(P\|Q) &= \sup_g \left\{ \mathbb{E}_P\left[g(x)\right] - \mathbb{E}_Q\left[e^{g(x)-1}\right] \right\} \\ &= \sup_g \sup_c \left\{ \mathbb{E}_P\left[(g+c)(x)\right] - \mathbb{E}_Q\left[e^{g(x)+c-1}\right] \right\} \\ &= \sup_g \left\{ \mathbb{E}_P\left[g\right] - \log \mathbb{E}_Q\left[e^g\right] \right\} \end{aligned}$$

This last expression is the Donsker-Varadhan representation of the Kullback-Leibler divergence ([DV83]).

2. For the chi-square divergence:

$$f_{\text{ext}}(x) = (x-1)^2$$
$$f^*_{\text{ext}}(y) = y + \frac{y^2}{4}$$

Then:

$$\begin{aligned} \chi^2(P\|Q) =& \sup_g \left\{ \mathbb{E}_P(f) - \mathbb{E}_Q(g) - \frac{1}{4}\mathbb{E}_Q(g^2) \right\} \\ &= \sup_g \left\{ \mathbb{E}_P(g) - \mathbb{E}_Q(g) - \frac{1}{4}\mathbb{V}_Q(g) \right\} \\ &= \sup_g \sup_\lambda \left\{ \lambda\left[\mathbb{E}_p(g) - \mathbb{E}(g)\right] - \frac{1}{4}\lambda^2\mathbb{V}_Q(g) \right\} \end{aligned}$$

To conclude:

$$\chi^2(P\|Q) = \sup_g \frac{(\mathbb{E}_P g - \mathbb{E}_Q g)^2}{\mathbb{V}_Q(g)}$$

The chi-square divergence is special because most f-divergence are "locally chi-square". The following theorem precises what this last statement means:

Theorem 6. *Let f be a twice continuously differentiable convex function such that $\limsup_{x\to+\infty} f''(\lambda) < +\infty$. Then:*

1. *if $\chi^2(P\|Q) < +\infty$ then for any $0 < \lambda < 1$:*

$$D_f(\lambda P + \overline{\lambda} Q\|Q) < +\infty$$

2. *We have*

$$\lim_{\lambda\to 0}\frac{1}{\lambda^2}D_f(\lambda P+\overline{\lambda}Q\|Q)=\frac{1}{2}f''(1)\chi^2(P\|Q) \tag{1}$$

where the right-hand side is infinite if $\chi^2(P\|Q)=\infty$ *and* $f''(1)>0$*).*

***Remark* 4 -** a way to remember this last theorem : when λ goes to 0, we have that $\lambda P+\overline{\lambda}Q$ goes to Q. For $P\to Q$, we obtain the quadratic approximation:

$$D_f(P\|Q)=f''(1)\chi^2(P\|Q)(1+o(1))$$

Proof. 1. We have:

$$f(1+u)=f(1)+uf'(1)+u^2\int_0^1(1-\sigma)f''(1+u\sigma)\,\mathrm{d}\sigma$$

WLOG we assume f(1)=f'(1)=0. Then:

$$\begin{aligned}D_f(\lambda P+\overline{\lambda}Q\|Q)&=\int \mathrm{d}Qf\left(1+\lambda\frac{\mathrm{d}P-\mathrm{d}Q}{\mathrm{d}Q}\right)\\&=\int \mathrm{d}Q\left(\lambda\frac{\mathrm{d}P-\mathrm{d}Q}{\mathrm{d}Q}\right)^2\int_0^1\mathrm{d}\sigma(1-\sigma)f''\left(1+\sigma\lambda\frac{\mathrm{d}P-\mathrm{d}Q}{\mathrm{d}Q}\right)\end{aligned}$$

Since $f''>0$ (f convex) and since $1+\sigma\lambda\frac{\mathrm{d}P-\mathrm{d}Q}{\mathrm{d}Q}\geq 1-\lambda$, we obtain:

$$D_f(\lambda P+\overline{\lambda}Q\|Q)\leq\frac{1}{2}C_\lambda\lambda^2\chi^2(P\|Q)$$

2. The last inequality implies that if $\chi^2(P\|Q)<+\infty$, the dominated convergence theorem applies:

$$\begin{aligned}\frac{1}{\lambda^2}D_f(\lambda P+\overline{\lambda}Q\|Q)&=\int \mathrm{d}Q\left(\frac{\mathrm{d}P-\mathrm{d}Q}{\mathrm{d}Q}\right)^2\underbrace{f''\left(1+\sigma\lambda\frac{\mathrm{d}P-\mathrm{d}Q}{\mathrm{d}Q}\right)}_{\to f''(1)}\times\underbrace{\int_0^1(1-\sigma)\,\mathrm{d}\sigma}_{=1/2}\\&\longrightarrow\frac{1}{2}\chi^2(P\|Q)f''(1),\ \lambda\to 0\end{aligned}$$

We proved the case $\chi^2(P\|Q)<+\infty$. The case $\chi^2(P\|Q)=+\infty$ follows immediately (?).

□

I Application: Empirical distribution and χ^2-information

Consider an arbitrary channel $P_{Y|X}$ and some input distribution P_X. Suppose that we have $X_i \stackrel{iid}{\sim} P_X$ for $i = 1, \ldots, n$. Let

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

denote the empirical distribution corresponding to this sample. Let $P_Y = P_{Y|X} \circ P_X$ be the output distribution corresponding to P_X and $P_{Y|X} \circ \hat{P}_n$ be the output distribution corresponding to $\hat{P}_n$ (a random distribution). Note that when $P_{Y|X=x}(\cdot) = \phi(\cdot - x)$, where ϕ is a fixed density, we can think of $P_{Y|X} \circ \hat{P}_n$ as a *kernel density estimator (KDE)*, whose density is $\hat{p}_n(x) = (\phi * \hat{P}_n)(x)\frac{1}{n}\sum_{i=1}^{n} \phi(X_i - x)$. Furthermore, using the fact that $\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n] = P_Y$, we have

$$\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_X)] = D(P_Y \| P_X) + \mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)]\,,$$

where the first term represents the bias of the KDE due to convolution and increases with bandwidth of ϕ, while the second term represents the variability of the KDE and decreases with the bandwidth of ϕ. Surprisingly, the second term is is sharply (within a factor of two) given by the I_{χ^2} information. More exactly, we prove the following result.

Proposition 1. *We have*

$$\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] \le \log\left(1 + \frac{1}{n} I_{\chi^2}(X;Y)\right)\,, \tag{2}$$

where $I_{\chi^2}(X;Y) \triangleq \chi^2(P_{X,Y} \| P_X P_Y)$. *Furthermore,*

$$\liminf_{n\to\infty} n\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] \ge \frac{\log e}{2} I_{\chi^2}(X;Y)\,. \tag{3}$$

In particular, $\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] = O(1/n)$ *if* $I_{\chi^2}(X;Y) < \infty$ *and* $\omega(1/n)$ *otherwise.*

Proof. First, a simple calculation shows that

$$\mathbb{E}[\chi^2(P_{Y|X} \circ \hat{P}_n \| P_Y)] = \frac{1}{n} I_{\chi^2}(X;Y)\,.$$

Then from (**??**) and Jensen's inequality we get (2).

To get the lower bound in (3), let $\bar{X}$ be drawn uniformly at random from the sample $\{X_1, \ldots, X_n\}$ and let $\bar{Y}$ be the output of the $P_{Y|X}$ channel with input $\bar{X}$. With this definition we have:

$$\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] = I(X^n; \bar{Y})\,.$$

Next, apply (**??**) to get

$$I(X^n; \bar{Y}) \ge \sum_{i=1}^{n} I(X_i; \bar{Y}) = nI(X_1; \bar{Y})\,.$$

Finally, notice that

$$I(X_1;\bar{Y}) = D\left(\frac{n-1}{n}P_X P_Y + \frac{1}{n}P_{XY}\middle\|P_X P_Y\right)$$

and apply the local expansion of KL divergence (1) to get (3). ☐

In the discrete case, by taking $P_{Y|X}$ to be the identity ($Y = X$) we obtain the following guarantee on the closeness between the empirical and the population distribution. This fact can be used to test whether the sample was truly generated by the distribution P_X.

Corollary 2. *Suppose P_X is discrete with support $\mathscr{X}$, If $\mathscr{X}$ is infinite, then*

$$\lim_{n\to\infty} n\mathbb{E}[D(\hat{P}_n\|P_X)] = \infty\,. \tag{4}$$

Otherwise, we have

$$\mathbb{E}[D(\hat{P}_n\|P_X)] \le \frac{\log e}{n}(|\mathscr{X}|-1)\,. \tag{5}$$

Proof. Simply notice that $I_{\chi^2}(X;X) = |\mathscr{X}|-1$. ☐

Application to KDE:

Let $\phi_\varepsilon = \mathscr{N}(0,\varepsilon)$ and choose

$$\begin{cases} P_{Y|X=x} = \mathscr{N}(x,\varepsilon) \\ \tilde{P}_{n,\varepsilon} := P_{Y|X}\circ\hat{P}_n = \hat{P}_n * \phi_\varepsilon \end{cases}$$

We have:

$$\mathbb{E}\left[D(\tilde{P}_{n,\varepsilon}\|P*\phi_\varepsilon)\right] \asymp \frac{1}{n}I_{\chi^2}(X, X+\sqrt{\varepsilon}Z)$$

Since:

$$\mathbb{E}\left[D(\tilde{P}_{n,\varepsilon}\|P)\right] = \mathbb{E}\left[D(\tilde{P}_{n,\varepsilon}\|P*\phi_\varepsilon\right] + D(P*\phi_\varepsilon\|P)$$

Under smoothness assumption:

$$\begin{aligned} &I_{\chi^2}(X;X+\sqrt{\varepsilon}Z) \sim 1/\varepsilon \\ &D(P*\phi_\varepsilon\|P) = (\varepsilon + o(\varepsilon))I_F(P) \sim \varepsilon \\ &\mathbb{E}\left[D(\tilde{P}_{n,\varepsilon}\|P)\right] \asymp \frac{1}{n\varepsilon} + \varepsilon \end{aligned}$$

Which implies:

$$\inf_\varepsilon \mathbb{E}\left[D(\tilde{P}_{n,\varepsilon}\|P)\right] \preceq \frac{1}{\sqrt{n}}$$

Theorem 7 (Hammersley-Chapman-Robbins bound [Ham50], [CR+51]). *For all $\hat{\theta}, \theta_1, \theta_2$ in $\mathbb{R}$:*

$$\mathbb{E}^{\theta_1}\left[(\hat{\theta}-\theta_1)^2\right] \geq \frac{\left[\mathbb{E}^{\theta_1}(\hat{\theta})-\mathbb{E}^{\theta_2}(\hat{\theta})\right]^2}{\chi^2(P^{\theta_2}\|P^{\theta_1})}$$

Proof. This last statement is simply the application of an earlier result:

$$\chi^2(P^{\theta_2}\|P^{\theta_1}) \geq \frac{\left[\mathbb{E}^{\theta_1}(\hat{\theta}-\theta_1)-\mathbb{E}^{\theta_2}(\hat{\theta}-\theta_1)\right]^2}{\mathbb{V}_{\theta_1}(\hat{\theta}-\theta_1)}$$

□

Theorem 8 (f-divergences are locally Fisher info). *Under regularity condition on $\{P^\theta\}$ we have*

$$\chi^2(P^{\theta_1}\|P^{\theta_2}) = (\theta_1-\theta_2)^2 I_F(\theta_2) + o((\theta_1-\theta_2)^2)$$
$$D_f(P^{\theta_1}\|P^{\theta_2}) = \frac{1}{2}f''(1)(\theta_1-\theta_2)^2 I_F(\theta_2) + o((\theta_1-\theta_2)^2)$$

Here, we suppose that $\mathbb{E}^\theta(\hat{\theta}) = \theta$ i.e. that $\hat{\theta}$ is unbiased.

Corollary 3 (Cramer-Rao). *Supposing that $\hat{\theta}$ is unbiased:*

$$\mathbb{E}^{\theta_1}\left[(\hat{\theta}-\theta_1)^2\right] \geq \lim_{\theta_2\to\theta_1} \frac{(\theta_2-\theta_1)^2}{\chi^2\left(P^{\theta_2}\|P^{\theta_1}\right)} = \frac{1}{I_F(\theta_1)}$$

Corollary 4 (Biased Cramer-Rao). *Denoting by $b(\theta) = \mathbb{E}^\theta(\hat{\theta}) - \theta$:*

$$\mathbb{E}^{\theta_1}\left[(\hat{\theta}-\theta_1)^2\right] \geq b(\theta_1)^2 + \frac{1+b'(\theta_1)^2}{I_F(\theta_1)}$$

Theorem 9 (Van Trees [Tre68]). *Let π be a density on Θ. Then:*

$$\mathbb{E}_{\theta\sim\pi}\mathbb{E}^\theta_{X_1^n \overset{i.i.d.}{\sim} P^\theta}\left[(\hat{\theta}-\theta)^2\right] \geq \frac{1}{I_F(\pi)+\mathbb{E}_{\theta\sim\pi}\left[I_F(\theta)\right]}$$

where $I_F(\pi) := \int \frac{\pi'^2}{\pi}$.

Corollary 5. *Under regularity assumptions:*

$$R_n^* = \frac{1+o(1)}{n\inf_{\theta\in\Theta} I_F(\theta)}$$

"Nice" proof of Van Trees' inequality. Let R_δ be the distance $\pi(\cdot-\delta)$.

$$P_{\theta,X} : \begin{cases} \theta \sim R_\delta \\ X \sim P^{\theta-\delta} \end{cases} \qquad Q_{\theta,X} : \begin{cases} \theta \sim R_0 \\ X \sim P^{\theta} \end{cases}$$

Note that $P_X = Q_X$. From variational characterization we get:

$$\mathbb{V}_Q(\theta - \hat{\theta}) \geq \frac{\left(\mathbb{E}_Q\left[\hat{\theta} - \theta\right] - \mathbb{E}_p\left[\hat{\theta} - \theta\right]\right)^2}{\chi^2(P_{\theta,X} \| Q_{\theta,X})}$$

under both Q and P, $\hat{\theta}$ has the exactly the same distribution. The last inequality yields:

$$\mathbb{V}_Q(\theta - \hat{\theta}) \geq \frac{\delta^2}{\chi^2(P_{\theta,X} \| Q_{\theta,X})},\ \delta \to 0,\ p\theta - \delta \to p\theta$$

We simply apply Taylor-Young:

$$\begin{aligned} \chi^2(P_{\theta,X} \| Q_{\theta,X}) &= \underbrace{\chi^2(P_\theta \| Q_\theta)}_{\chi^2(R_\delta \| R_0)} + \mathbb{E}_{\theta\sim\pi}\left(\frac{P_\theta}{Q_\theta}\right)^2 \underbrace{\chi_2(P_{\theta-\delta} \| P_\theta)}_{\text{loc. Fisher information}} \\ &= \delta^2 I_F(\pi) + \delta^2 \mathbb{E}_\theta I_F(\theta) + o(\delta^2),\ \delta \to 0 \end{aligned}$$

□

This is the translation of Van Trees' inequality into "information-theoretic vocabulary". The advantage of the latter is that it can be applied also in cases where Fisher information does not exist or non-regular, and thus obtain rates other than $\frac{1}{n}$.

Bibliography

[CR+51] Douglas G Chapman, Herbert Robbins, et al. Minimum variance estimation without regularity assumptions. *The Annals of Mathematical Statistics*, 22(4):581–586, 1951.

[DV83] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

[GI59] Yaglom A.M. Gelfand I.M. Calculation of the amount of information about a random function obtained in another function'. *American Mathematical Society Translation Series*, 2:12, 1959.

[Ham50] John M Hammersley. On estimating restricted parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, 12(2):192–240, 1950.

[Per59] Albert Perez. Information theory with an abstract alphabet (generalized forms of mcmillan's limit theorem for the case of discrete and continuous times. *Theory of Probability & Its Applications*, 4(1):99–102, 1959.

[Tre68] Harry L. Van Trees. Detection, estimation, and modulation theory. I, 1968.