Lecture One - STAT 212a

Aditya Guntuboyina

August 29, 2012

1 \textit{$f$-divergences}

Let $f: (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. By virtue of convexity, both the limits $f(0) := \lim_{x \downarrow 0} f(x)$ and $f'(\infty) := \lim_{x \uparrow \infty} f(x)/x$ exist, although they may be equal to $+\infty$. For two probability measures $P$ and $Q$, the $f$-divergence $D_f(P||Q)$ is defined by

$$D_f(P||Q) := \int \left\{ q > 0 \right\} f \left( \frac{p}{q} \right) dQ + f'(\infty) P \{ q = 0 \},$$

where $p$ and $q$ denote the densities of $P$ and $Q$ with respect to a common measure $\lambda$. The definition does not depend on the choice of the dominating measure $\lambda$.

Work out examples: KL, Total variation, Hellinger, chi-squared.

2 \textit{Representation via Testing}

Take two probability measures $P$ and $Q$. Given an observation $X$ consider the problem of testing $H_0 : X \sim P$ against $H_1 : X \sim Q$. A test would be a function of $X$ taking values in $\{0, 1\}$.

Any test $T$ would make two kinds of errors: $Q(T(X) = 0)$ and $P(T(X) = 1)$.

Let $R_w(T)$ denote the convex combination of these two errors:

$$R_w(T) := wQ(T(X) = 0) + (1 - w)P(T(X) = 1) \quad \text{for } w \in [0, 1].$$

The problem of finding $T$ so that $R_w(T)$ is minimized is easily solved. Let
$p$ and $q$ denote densities of $P$ and $Q$ with respect to $\lambda$. Write

$$R_w(T) = w \int \{T(x) = 0\} q(x) d\lambda(x) + (1 - w) \int \{T(x) = 1\} p(x) d\lambda(x)$$

$$= \int \{\{T(x) = 0\} wq(x) + \{T(x) = 1\}(1 - w)p(x)\} d\lambda(x)$$

$$\geq \int \min(wq(x), (1 - w)p(x)) d\lambda(x),$$

with equality when $T(X)$ is chosen as 0 when $wq(X) \geq (1 - w)p(X)$ and 1 otherwise. Therefore

$$L_{P,Q}(w) := \inf_T R_w(T) := \int \min(wq, (1 - w)p)$$

Note that the above argument would also hold if we allow tests to take arbitrary values in $[0, 1]$ (as opposed to just 0 or 1). Such tests are called randomized tests. In this case, the type 1 error would be defined as $\int T dP$ and the type 2 error is $\int (1 - T)dQ$ and we would have

$$\inf_{T:X \to [0,1]} \left( w \int (1 - T)dQ + (1 - w) \int TdP \right) = \int \min(wq, (1-w)p)d\lambda = L_{P,Q}(w).$$

$L(w)$ is the smallest average error in the testing problem $P$ against $Q$. As a function of $w$ on $[0, 1]$, it is clear that $L(w)$ is concave and that it satisfies $0 \leq L(w) \leq \min(w, 1 - w)$.

Moreover, it can be shown that for every concave function $u$ on $[0, 1]$ satisfying $0 \leq u(w) \leq \min(w, 1 - w)$, there exist two probability measures $P$ and $Q$ such that $u = L_{P,Q}$. In other words, convexity and lying inside the triangle uniquely characterize $L_{P,Q}$:

**Theorem 2.1.** Let $\phi$ be any concave function on $[0, 1]$ satisfying the inequality $0 \leq \phi(t) \leq \min(t, 1 - t)$ for all $0 \leq t \leq 1$. Then there exist two probability measures $P$ and $Q$ on $[0, 1]$ such that $\phi = \phi_{P,Q}$.

**Proof.** Let $\psi$ be the nonnegative function on $[0, \infty)$ defined by $\psi(s) := (1 + s)\phi(s/(1 + s))$ for $s \geq 0$. Because $\phi$ is concave and continuous on $[0, 1]$, there exists a countable collection of linear functions $t \mapsto d_i + c_i t$ such that $\phi(t) = \inf_{i \geq 1} (d_i + c_i t)$ for all $t \in [0, 1]$. It follows that $\psi(s) = \inf_{i \geq 1} (d_i + (c_i + d_i)s)$ for all $s \geq 0$ and hence $\psi$ is concave on $[0, \infty)$. The fact that $\phi(1) = 0$ implies that $\inf_{i \geq 1} (d_i + c_i) \geq 0$. Therefore, $\psi$ is a non-decreasing function on $[0, \infty)$. Finally, $\phi(t) \leq \min(t, 1 - t)$ yields $\psi(s) \leq \min(s, 1)$ for all $s \geq 0$.

These properties of the function $\psi$ allow us to represent $\psi(s)$ as $\int_0^s h(x)dx$ for a nonnegative, non-increasing, right continuous function $h$ on $[0, \infty)$ satisfying $h(0) \leq 1$ and $h(x) \downarrow 0$ as $x \to \infty$. If $h(0)$ equals 0, then $\psi$ and hence $\phi$
is identically equal to zero and we can then take $P$ and $Q$ to be supported on disjoint subintervals of $[0, 1]$. Thus we shall assume that $h(0) > 0$. For $0 < y < h(0)$, let us define

$$h^-(y) := \inf \{ x \geq 0 : h(x) \leq y \}.$$ 

Since $h$ is right continuous and non-increasing, it can be easily be checked that

$$h(x) > y \text{ if and only if } x < h^-(y).$$

We can now define $P$ and $Q$ for which $\phi_{P,Q} = \phi$. Let $\lambda = m + \delta_1$, where $m$ denotes the Lebesgue measure on $[0, 1]$ and $\delta_1$ is the Dirac measure concentrated at the point 1. The probability measures $P$ and $Q$ have densities $p$ and $q$ with respect to $\lambda$ that are given by

$$p(y) := h^{-1}(y) \{ 0 < y < h(0) \} + (1 - \psi(\infty)) \{ y = 1 \} \text{ and } q(y) = \{ 0 \leq y < 1 \},$$

where $\psi(\infty) := \lim_{s \to \infty} \psi(s)$. Note that $Q$ equals $m$. Observe that $P$ is indeed a probability measure because

$$\int_0^{h(0)} h^-(y) dy + \int_0^1 \int_0^\infty \{ x < h^-(y) \} \{ y \leq h(0) \} dxdy = \int_0^\infty h(x) dx = \psi(\infty).$$

We shall now show that the function $\psi_{P,Q}$ defined by $\psi_{P,Q}(s) := \int \min(p, sq)d\lambda$ on $[0, \infty)$ equals the function $\psi$. It should be obvious that this implies that $\phi_{P,Q} = \phi$. For a fixed $s \geq 0$, we can write

$$\psi_{P,Q}(s) = \int_0^1 \min(p(y), sq(y))d\lambda(y)$$

$$= \int_0^1 \min(h^-(y), s) \{ 0 < y < h(0) \} dy$$

$$= \int_0^1 \int_0^\infty \{ x < s \} \{ x < h^-(y) \} \{ 0 < y < h(0) \} dxdy$$

$$= \int_0^\infty \int_0^1 \{ x < s \} \{ y < h(x) \} \{ 0 < y < h(0) \} dydx$$

$$= \int_0^s h(x) dx = \psi(s).$$

The proof is complete. $\square$

$L$ denotes the smallest average error in hypothesis testing. If $P$ and $Q$ are close, then it is hard to test between them so $L$ will be high. On the other hand, if $P$ and $Q$ are far, then testing is easy so $L$ will be small. Indeed, when
\( P = Q \), the function \( L_{P,Q} \) equals \( \min(w, 1 - w) \). On the other hand, when \( P \) and \( Q \) are mutually singular, \( L_{P,Q} \) equals 0. Thus, the difference between the functions \( \min(w, 1 - w) \) and \( L_{P,Q} \) is a measure of the distance between \( P \) and \( Q \): the function:

\[
    w \mapsto \min(w, 1 - w) - L_{P,Q}(w) \quad \text{for } w \in [0, 1]
\]
is a reasonable measure of distance between the probability measures \( P \) and \( Q \).

But it is for many purposes desirable to have real-valued measures of distance as opposed to functions. A natural idea is to integrate the above function with respect to a measure on \((0, 1)\). Therefore, for every measure \( \nu \) (not necessarily finite) on \((0, 1)\), we define

\[
    I_{\nu}(P||Q) := \int (\min(w, 1 - w) - L_{P,Q}(w)) \, d\nu(w).
\]

It turns out that these resulting measures of distance are precisely the same as \( f \)-divergences. In other words, given every \( f \)-divergence \( D_f \), there exists a measure \( \nu_f \) on \((0, 1)\) such that \( D_f \) is the same as \( I_{\nu_f} \).

(As an example, let \( \nu \) denote the point mass at \( w = 1/2 \). Then it is easy to check that \( I_{\nu} \) equals \( V/2 \) where \( V \) denotes the total variation distance.)

Here is the proof of this fact. Let \( f : (0, \infty) \to \mathbb{R} \) be a convex function with \( f(1) = 0 \). At every point in \((0, \infty)\), the function \( f \) has a finite right derivative \( f'_r \) that is non-decreasing, right continuous and satisfies

\[
    f(b) - f(a) = \int_a^b f'_r(y) \, dy \quad \text{for } 0 < a < b < \infty.
\]

Observe that the function \( f'_r \) defines a measure \( \mu_f \) on \((0, \infty)\) by

\[
    \mu_f(a, b] = f'_r(b) - f'_r(a) \quad \text{for } 0 < a < b < \infty.
\]

Lemma 2.2. For each \( x \in (0, \infty) \), we have

\[
    f(x) = f'_r(1)(x - 1) + \int_0^x (x<s<1) + s\{s \leq 1\} - \min(x,s) \, d\mu_f(s). \tag{1}
\]

Further, the limits \( f(0) := \lim_{x \downarrow 0} f(x) \) and \( f'(\infty) := \lim_{x \uparrow \infty} f(x)/x \) both exist and satisfy

\[
    f(0) = \mu_f(\{s \leq 1\}) - f'_r(1) \quad \text{and} \quad f'(\infty) = \mu_f(1, \infty) + f'_r(1).
\]
Proof. When \( x > 1 \), we can write

\[
    f(x) = f(x) - f(1) \\
    = \int_1^x f'_r(y)dy \\
    = f'_r(1)(x-1) + \int_1^x (f'_r(y) - f'_r(1)) dy \\
    = f'_r(1)(x-1) + \int_1^x \mu_f(1,y)dy \\
    = f'_r(1)(x-1) + \int_1^x \int_y^x d\mu_f(s)dy \\
    = f'_r(1)(x-1) + \int \{1 < y \leq x\} dy d\mu_f(s) \\
    = f'_r(1)(x-1) + \int_{-\infty}^x (x-s)^+ d\mu_f(s).
\]

Similarly for \( 0 < x < 1 \), we have

\[
    f(x) = -(f(1) - f(x)) \\
    = -\int_x^1 f'_r(y)dy \\
    = f'_r(1)(x-1) + \int_x^1 (f'_r(1) - f'_r(y)) dy \\
    = f'_r(1)(x-1) + \int_x^1 \mu_f(y,1)dy \\
    = f'_r(1)(x-1) + \int_x^1 \int_y^1 d\mu_f(s)dy \\
    = f'_r(1)(x-1) + \int \{x \leq y < 1\} dy d\mu_f(s) \\
    = f'_r(1)(x-1) + \int_0^1 (s-x)^+ d\mu_f(s).
\]

Both these two expressions for \( x > 1 \) and \( 0 < x < 1 \) can be simultaneously expressed as

\[
    f(x) = f'_r(1)(x-1) + \int ((x-s)^+ \{s > 1\} + (s-x)^+ \{s \leq 1\}) d\mu_f(s).
\]

Writing \((a-b)^+\) as \(a - \min(a,b)\), we get (1). Also

\[
    f(0) = \lim_{x \downarrow 0} f(x) = \lim_{x \downarrow 0} \int (s-x)^+ \{s \leq 1\} d\mu_f(s) - f'_r(1) \\
    = \int s \{s \leq 1\} d\mu_f(s) - f'_r(1),
\]

5
where the last equality follows by monotone convergence. Finally,
\[
f'(\infty) = \lim_{x \uparrow \infty} \frac{f(x)}{x} = f'_r(1) + \lim_{x \uparrow \infty} \int \left(1 - \frac{s}{x}\right)^+ \{s > 1\} d\mu_f(s)
\]
\[
= f'_r(1) + \mu_f(1, \infty),
\]
where we have again used monotone convergence. 

We can now prove the connection between \( f \)-divergences and testing.

**Theorem 2.3.** For every \( f \)-divergence \( D_f \), there exists a measure \( \nu_f \) on \((0, 1)\) such that \( D_f(P||Q) = I_{\nu_f}(P||Q) \) for all probability measures \( P \) and \( Q \). In particular, when \( f \) is smooth, we have the following representation for \( D_f(P||Q) \):

\[
D_f(P||Q) = \int_0^1 \left( \min(1 - w, w) - \int \min((1 - w)p, wq) \right) f''(\frac{w}{1-w}) \frac{dw}{(1-w)^3}.
\]

**Proof.** We start with
\[
D_f(P||Q) := \int \left[ \{q > 0\} f(\frac{p}{q}) \right] dQ + f'(\infty) P\{q = 0\}.
\]
Using the formula for \( f(p/q) \) from Lemma 2.2, we get that the term
\[
\int \left[ f'_r(1) \left( \frac{p}{q} - 1 \right) + \int \left( \frac{p}{q} \{s > 1\} + s\{s \leq 1\} - \min(p/q, s) \right) d\mu_f(s) \right] q\{q > 0\} d\lambda
\]
which is the same as
\[
f'_r(1) (P\{q > 0\} - 1) + \int \left[ p\{s > 1\} + qs\{s \leq 1\} - \min(p, qs) \right] q\{q > 0\} d\lambda d\mu_f(s).
\]
and
\[
f'_r(1) (P\{q > 0\} - 1) + \int \left[ \{s > 1\} P\{q > 0\} + s\{s \leq 1\} - \int \min(p, qs) d\lambda \right] d\mu_f(s).
\]
Again, using Lemma 2.2 for the expression for \( f'(\infty) \), we obtain
\[
f'(\infty) P\{q = 0\} = \mu_f(1, \infty) + f'_r(1) P\{q = 0\}
\]
\[
= f'_r(1) P\{q = 0\} + \int \{s > 1\} P\{q = 0\} d\mu_f(s).
\]
Putting these together, we get that \( D_f(P||Q) \) equals
\[
\int \left[ \{s > 1\} + s\{s \leq 1\} - \int \min(p, qs) d\lambda \right] d\mu_f(s),
\]

6
which can be written as

\[ D_f(P||Q) = \int \left[ \min(1, s) - \int \min(p, qs) d\lambda \right] d\mu_f(s). \]

We now do the change variable \( s = w/(1 - w) \) or \( w = s/(1 + s) \). The theorem will then be proved with \( \nu_f \) being the image of \( \mu_f \) under the transformation \( s \mapsto s/(1+s) \). In the special case when \( f \) is smooth, \( d\mu_f(s) = f''(s)ds \) and the change of variable can be carried out explicitly resulting in the expression (2.3).

3 Data-Processing Inequality

The data-processing inequality is one important property shared by all \( f \)-divergences. It is easily proved through Theorem 2.3. There are two versions of this: the simple version and the more general version. In the following, we are ignoring measurability issues.

Let us start with the simpler version. Let \( P \) and \( Q \) be two probability measures on a space \( \mathcal{X} \). Let \( \Gamma \) be a mapping from \( \mathcal{X} \) to another space \( \mathcal{Y} \). Consider now the probability measures \( P_{\Gamma^{-1}} \) and \( Q_{\Gamma^{-1}} \) on the space \( \mathcal{Y} \) (If \( P \) is the distribution of a random object \( X \), then \( P_{\Gamma^{-1}} \) is the distribution of the random object \( \Gamma(X) \)). The data-processing inequality says that \( D_f(P_{\Gamma^{-1}}||Q_{\Gamma^{-1}}) \leq D_f(P||Q) \). The proof is almost immediate from Theorem 2.3. Any test for testing between the hypotheses \( P_{\Gamma^{-1}} \) against \( Q_{\Gamma^{-1}} \) is automatically also a test for \( P \) against \( Q \). It follows, therefore, that \( L_{P,Q}(w) \leq L_{P_{\Gamma^{-1},Q_{\Gamma^{-1}}}}(w) \) for all \( w \) thereby completing the proof.

The more general version of the data processing theorem involves randomization or kernels. A kernel \( K \) from a space \( \mathcal{X} \) to another measurable space \( (\mathcal{Y}, B) \) is a function defined on \( \mathcal{X} \times B \) such that \( K(x, \cdot) \) is a probability measure on \( \mathcal{Y} \) for each \( x \in \mathcal{X} \). For every probability measure \( P \) on \( \mathcal{X} \), one can define a probability measure \( KP \) on \( \mathcal{Y} \) by the formula:

\[ KP(B) = \int K(x, B) dP(x). \]

Another way of writing this would be

\[ d(KP)(y) = \int_X K(x, dy) dP(x). \]

The data processing theorem then states that \( D_f(KP||KQ) \geq D_f(P||Q) \). When there is a mapping \( \Gamma \) from \( \mathcal{X} \) to \( \mathcal{Y} \), then one can define the kernel \( K(x, \cdot) \) as the point probability measure at \( T(x) \) and then this version yields the previous version of the data processing theorem.
To prove this general version, we use

\[ L_{KP,KQ}(w) = \inf_{T: \mathcal{Y} \to [0,1]} \left( w \int (1 - T)d(KQ) + (1 - w) \int Td(KP) \right) \]

For every \( T: \mathcal{Y} \to [0,1] \) (i.e., \( T \) is a randomized test for testing \( KP \) against \( KQ \)), define

\[ \tilde{T}(x) := \int_{\mathcal{Y}} T(y)K(x,dy). \]

Clearly \( \tilde{T} \) is a randomized test for testing \( P \) against \( Q \). Now,

\[ \int_{\mathcal{X}} \tilde{T}(x)dP(x) = \int_{\mathcal{X}} \int_{\mathcal{Y}} T(y)K(x,dy)dP(x) = \int_{\mathcal{Y}} T(y) \int_{\mathcal{X}} K(x,dy)dP(x) = \int_{\mathcal{Y}} Td(KP), \]

and similarly \( \int (1 - \tilde{T})dQ = \int (1 - T)d(KQ) \). Thus

\[ L_{P,Q}(w) \leq w \int (1 - \tilde{T})dQ + (1 - w) \int \tilde{T}dP \]

\[ = w \int (1 - T)d(KQ) + (1 - w) \int Td(KP). \]

This inequality is true for all \( T: \mathcal{Y} \to [0,1] \) which implies that \( L_{P,Q}(w) \leq L_{KP,KQ}(w) \) for all \( w \). Theorem 2.3 now gives \( D_f(P||Q) \geq D_f(KP||KQ) \).

In the next class, we will apply Theorem 2.3 to obtain inequalities between \( f \)-divergences.