Lecture-23: Variational representation of divergences

1 Variational representation of f-divergences

Theorem 1.1. *Let $X : \Omega \to \mathcal{X}$ be a random variable on $(\Omega, \mathcal{F})$ and $P, Q \in \mathcal{M}(\mathcal{X})$. Given finite partition $\mathcal{E} \triangleq \{E_1, \dots, E_n\} \subseteq \mathcal{F}$ of Ω, we define the distribution $P_{\mathcal{E}} \in \mathcal{M}([n])$ by $P_{\mathcal{E}}(i) \triangleq P(E_i)$ and $Q_{\mathcal{E}}(i) \triangleq Q(E_i)$ for all $i \in [n]$. Then*

$$D_f(P\|Q) = \sup_{\mathcal{E} \subseteq \mathcal{F} : \mathcal{E} \text{ finite partition of } \Omega} D_f(P_{\mathcal{E}} \| Q_{\mathcal{E}}).$$

Proof. Let $P, Q \in \mathcal{M}(\mathcal{X})$. If $P \not\ll Q$, then $D_f(P\|Q) = \infty$, and there exists $E \in \mathcal{F}$ such that $P(E) > 0 = Q(E)$. For finite partition $\mathcal{E} \triangleq (E, E^c)$, we have $D_f(P_{\mathcal{E}}\|Q_{\mathcal{E}}) = \infty$.

Therefore, we can assume that $P \ll Q$ and relative density $g \triangleq \frac{dP}{dQ}$ exists. Let $X_n : \Omega \to \mathcal{X}$ be a simple function such that $X_n \leqslant X$ and $\mathcal{E} \triangleq \{X_n^{-1}\{x\} : x \in \mathcal{X}\}$, then $P_{\mathcal{E}} \triangleq P_{X_n}$ and $Q_{\mathcal{E}} \triangleq Q_{X_n}$. We can consider a Markov chain $X \to X_n$ and it follows from data processing inequality for f divergences that $D_f(P\|Q) \geqslant D_f(P_{\mathcal{E}_n}\|Q_{\mathcal{E}_n})$.

We first assume that $g \geqslant 0$ and define $E_n \triangleq (g \circ X)^{-1}[\epsilon(n-1), \epsilon n)$ for each $n \in \mathbb{N}$ and a fixed $\epsilon > 0$. Defining $S \triangleq \{\omega \in \Omega : (g \circ X)(\omega) > 0\}$ and $E_\infty \triangleq \{\omega : (g \circ X)(\omega) = 0\}$, it follows that

$$\epsilon \sum_{n \in \mathbb{N}} (n-1) Q(E_n) \leqslant \int_S dQ f(\frac{dP}{dQ}) \leqslant \epsilon \sum_{n \in \mathbb{N}} n Q(E_n) + f(0) Q(E_\infty) \leqslant \epsilon \sum_{n \in \mathbb{N}} (n-1) Q(E_n) + f(0) Q(E_\infty) + \epsilon.$$

□

Definition 1.2 (convex conjugate). Let $f : (0, \infty) \to \mathbb{R}$ be a convex function, then its *convex conjugate* $f^* : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is defined as $f^*(y) \triangleq \sup_{x \in \mathbb{R}_+} xy - f(x)$ for all $y \in \mathbb{R}$. The domain of convex conjugate f^* is denoted by $\mathrm{dom}(f^*) \triangleq \{y \in \mathbb{R} : f^*(y) < \infty\}$.

Definition 1.3 (lower semicontinuity). A function $f : \mathcal{X} \to \mathbb{R}$ is *lower semicontinuous* if its epigraph $\{(x, t) \in \mathcal{X} \times \mathbb{R} : t \geqslant f(x)\}$ is closed in $\mathcal{X} \times \mathbb{R}$.

Lemma 1.4. *Consider a map $f : (0, \infty) \to \mathbb{R}$, then its convex conjugate $f^* : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ has the following two properties.*

*(a) **Convexity.** f^* is a convex map.*

*(b) **Biconjugation.** $f^{**} \leqslant f$ with equality iff f is convex and lower semicontinuous.*

Proof. Recall that f^* is the convex conjugate of f.

(a) We observe that $xy - f(x)$ is an affine map in y for each $x \in \mathbb{R}_+$, and since the supremum of affine maps is convex, it follows that f^* is convex in y.

(b) We will show the upper bound and the equality.

  (i) From definition of convex conjugate, we have $f(x) \geqslant xy - f^*(y)$ for each $x, y \in (\mathbb{R}_+ \times \mathbb{R})$. It follows that $f^{**}(x) \leqslant f(x)$ for all $x \in \mathbb{R}_+$.

  (ii) When f is convex and lower semicontinuous, its epigraph is convex and closed. Let $s \in \mathbb{R}_+$ such that $f^{**}(s) < f(s)$. It follows that $(s, f^{**}(s))$ is not in the epigraph of f. From separating hyperplane theorem, there exists a hyperplane $(y, -\lambda) \in \mathbb{R}^2$ and a scalar α such that $sy - \lambda f^{**}(s) > \alpha$ and $xy - \lambda f(x) \leqslant \alpha$ for each $x \in \mathbb{R}_+$. If $\lambda \neq 0$, then we get a contradiction, we can take $\lambda = 1$ without any loss of generality. We obtain that $f^*(y) \leqslant \alpha$ and $sy - f^*(y) > f^{**}(s)$, which is a contradiction.

□

Definition 1.5. Consider input space $\mathcal{X}$ and observation $X : \Omega \to \mathcal{X}$, then for any convex functional $\Psi : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$, we denote its associated convex conjugate as $\Psi^* : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$, defined for each map $g \in \mathbb{R}^{\mathcal{X}}$ as $\Psi^*(g) \triangleq \sup_P \mathbb{E}_{X \sim P} g(X) - \Psi(P)$.

Under appropriate conditions e.g. finite $\mathcal{X}$, biconjugation yields the sought-after variational representation $\Psi(P) = \sup_g \mathbb{E}_{X\sim P} g(X) - \Psi^*(g)$. We will now compute these conjugates for $\Psi(P) \triangleq D_f(P\|Q)$. It turns out to be convenient to first extend the definition of $D_f(P\|Q)$ to all finite signed measures P then compute the conjugate.

Definition 1.6. Let $f : (0, \infty) \to \mathbb{R}$ be a convex function, then we can define its convex extension as $f_{\text{ext}} : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ such that $f_{\text{ext}}(x) \triangleq f(x)$ for $x \in \mathbb{R}_+$ and f_{ext} is convex on $\mathbb{R}$.

Remark 1. In general, we can always choose $f_{\text{ext}}(x) = \infty$ for all $x < 0$. In special cases, e.g. $f(x) = \frac{|x-1|}{2}$ or $f(x) = (x-1)^2$ we can directly take $f_{\text{ext}}(x) = f(x)$ for all x.

Theorem 1.7. *Consider a random variable $X : \Omega \to \mathcal{X}$ on $(\Omega, \mathcal{F})$ and $P, Q, \mu \in \mathcal{M}(\mathcal{X})$ such that $P, Q \ll \mu$. Consider a convex function $f : (0, \infty) \to \mathbb{R}$, its extension $f_{\text{ext}} : \mathbb{R} \to \mathbb{R}$ and its convex conjugate $f^*_{\text{ext}} : \mathbb{R} \to \mathbb{R}$. Then,*

$$D_f(P\|Q) = \sup_{g:\mathcal{X}\to\text{dom}(f^*_{\text{ext}})} \mathbb{E}_{X\sim P} g(X) - \mathbb{E}_{X\sim Q} f^*_{\text{ext}}(g(X)), \tag{1}$$

where the supremum can be taken over either (a) all simple g or (b) over all g satisfying $\mathbb{E}_{X\sim Q} f^*_{\text{ext}}(g(X)) < \infty$.

Proof. The convex conjugate of convex extension f_{ext} is defined as $f^*_{\text{ext}}(y) \triangleq \sup_{x\in\mathbb{R}} xy - f_{\text{ext}}(x)$ for each $y \in \text{dom}(f^*_{\text{ext}})$ and we denote the relative densities of measures P, Q with respect to μ by p, q respectively.

Step 1. We show that for any $g : \mathcal{X} \to \text{dom}(f^*_{\text{ext}})$, we must have

$$\mathbb{E}_{X\sim P} g(X) \leqslant D_f(P\|Q) + \mathbb{E}_{X\sim Q} f^*_{\text{ext}}(g(X)). \tag{2}$$

Let $g : \mathcal{X} \to \text{dom}(f^*_{\text{ext}})$. From the definition of f^*_{ext} we have for every $x \in S \triangleq \{z \in \mathcal{X} : q(z) > 0\}$,

$$f^*_{\text{ext}}(g(x)) + f_{\text{ext}}\Big(\frac{p(x)}{q(x)}\Big) \geqslant g(x)\frac{p(x)}{q(x)}.$$

Recall that $D_{f_{\text{ext}}}(P\|Q) = \int_{x\in S} d\mu(x) q(x) f_{\text{ext}}\Big(\frac{p(x)}{q(x)}\Big) + f'_{\text{ext}}(\infty)\mathbb{E}_{X\sim P}\mathbb{1}_{\{q(X)=0\}}$. Multiplying both sides by $\mathbb{1}_S(x)$ and integrating both sides over $dQ = q d\mu$ restricted to the set $S \subseteq \mathcal{X}$, we get

$$\mathbb{E}_{X\sim Q} f^*_{\text{ext}}(g(X)) + D_{f_{\text{ext}}}(P\|Q) - f'_{\text{ext}}(\infty)\mathbb{E}_{X\sim P}\mathbb{1}_{\{q(X)=0\}} \geqslant \mathbb{E}_{X\sim P} g(X)\mathbb{1}_{\{q(X)>0\}}. \tag{3}$$

Recall that f_{ext} is convex and lower semicontinuous and hence $f^{**}_{\text{ext}} = f_{\text{ext}}$ by biconjugation, and hence $\frac{1}{x} f_{\text{ext}}(x) = \sup_{y\in\text{dom}(f^*_{\text{ext}})} y - \frac{1}{x} f^*_{\text{ext}}(y)$ for each $x \in \mathbb{R}$. Taking limit on both sides for $x \to \infty$, using the definition of $f'_{\text{ext}}(\infty)$, and the fact that $f^*_{\text{ext}}(y) < \infty$ for each $y \in \text{dom}(f^*_{\text{ext}})$, we obtain

$$f'_{\text{ext}}(\infty) = \lim_{x\to\infty} \frac{1}{x} f_{\text{ext}}(x) = \sup \text{dom}(f^*_{\text{ext}}).$$

Further, we have $f^*(g(x))q(x) + p(x)\frac{q(x)}{p(x)} f_{\text{ext}}(\frac{p(x)}{q(x)}) \geqslant g(x)p(x)$. Multiplying both sides by $\mathbb{1}_{\{q(x)=0\}} = 1 - \mathbb{1}_S(x)$ and integrating both sides over x, we obtain

$$f'_{\text{ext}}(\infty)\mathbb{E}_{X\sim P}\mathbb{1}_{\{q(X)=0\}} \geqslant \mathbb{E}_{X\sim P} g(X)\mathbb{1}_{\{q(X)=0\}}.$$

Summing this inequality with (3) we obtain the desired result in (2).

Step 2. We consider finite $\mathcal{X}$, and let $S \triangleq \{x \in \mathcal{X} : Q(x) > 0\}$ denote the support of Q. We show the following statement which is equivalent to (1),

$$D_f(P\|Q) = \sup_{g:S\to\text{dom}(f^*_{\text{ext}})} \mathbb{E}_{X\sim P} g(X) - \mathbb{E}_{X\sim Q} f^*_{\text{ext}}(g(X)) + f'(\infty)P(S^c). \tag{4}$$

Defining functional $\Psi(P) \triangleq \sum_{x\in S} Q(x) f_{\text{ext}}\Big(\frac{P(x)}{Q(x)}\Big) = D_f(P\|Q)$ where P takes values over all signed measures on S, we have $D_f(P\|Q) = \Psi(P) + f'(\infty)P(S^c)$. Functional Ψ is a function of $P \in \mathcal{M}(\mathcal{X})$, where $P(S^c) = P(\mathcal{X}) - P(S)$, and hence P can be identified with a map $\mathbb{R}^{\mathcal{X}}$. The convex conjugate of Ψ is $\Psi^* : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ defined for any $g \in \mathbb{R}^{\mathcal{X}}$, as

$$\begin{aligned}
\Psi^*(g) &\overset{(a)}{=} \sup_P \sum_{x\in\mathcal{X}} P(x)g(x) - Q(x)\left\{\sup_{h:S\to\text{dom}(f^*_{\text{ext}})} \frac{P(x)}{Q(x)} h(x) - f^*_{\text{ext}}(h(x))\right\} \\
&\overset{(b)}{=} \sup_P \inf_{h:S\to\text{dom}(f^*_{\text{ext}})} \sum_x P(x)(g(x) - h(x)) + Q(x) f^*_{\text{ext}}(h(x) \\
&\overset{(c)}{=} \inf_{h:S\to\text{dom}(f^*_{\text{ext}})} \sup_P \sum_x P(x)(g(x) - h(x)) + Q(x) f^*_{\text{ext}}(h(x),
\end{aligned}$$

where (a) follows from the fact that f_{ext} is convex and lower semicontinuous and hence by bi-conjugation, we have $f_{\text{ext}}(P(x)/Q(x)) = \sup_{h:S\to\text{dom}(f^*_{\text{ext}})} h(x)P(x)/Q(x) - f^*_{\text{ext}}(h(x))$ for each $x \in S$, (b) follows from finiteness of $\mathcal{X}$ to exchange infimum and the sum, and (c) follows from the minimax theorem which applies due to finiteness of $\mathcal{X}$. Since P is a signed measure, it follows that if $g \in \text{dom}(f^*_{\text{ext}})^S$, then the minimization is achieved for $g = h$ and $\Psi^*(g) = \mathbb{E}_{X\sim Q} f^*_{\text{ext}}(g(X))$. If $g \notin \text{dom}(f^*_{\text{ext}})^S$, then one can put arbitrarily large mass at $x \notin \text{dom}(f^*_{\text{ext}})$ to obtain $\Psi^*(g) = \infty$. That is, we have

$$\Psi^*(g) = \mathbb{E}_{X\sim Q} f^*_{\text{ext}}(g(X))\mathbb{1}_{\{g\in\text{dom}(f^*_{\text{ext}})^S\}} + \infty\mathbb{1}_{\{g\notin\text{dom}(f^*_{\text{ext}})^S\}}.$$

Recall that if Ψ is convex and lower semicontinuous, then the convex conjugate of Ψ^* is Ψ itself. Applying the convex duality of convex conjugates yields the proof of the desired (4).

Step 3. We show that supremum in (1) over simple functions g does yield $D_f(P\|Q)$, so that inequality (2) is tight. We will show that for simple functions, it suffices to show (1) for finite observation space $\mathcal{X}$. Indeed, for general $\mathcal{X}$, given a finite partition $\mathcal{E} \triangleq \{E_1, \dots, E_n\}$ of $\mathcal{X}$, we say a function $g : \mathcal{X} \to \mathbb{R}$ is $\mathcal{E}$-measurable if g is constant on each $E_i \in \mathcal{E}$. Taking the supremum over all finite partitions $\mathcal{E}$, we get fromTheorem 1.1 that

$$\begin{aligned} D_f(P\|Q) = \sup_{\mathcal{E}} D_f(P_{\mathcal{E}}\|Q_{\mathcal{E}}) &= \sup_{\mathcal{E}} \sup_{g\in\text{dom}(f^*_{\text{ext}})^{\mathcal{X}},\mathcal{E}-\text{measurable}} \mathbb{E}_{X\sim P} g(X) - \mathbb{E}_{X\sim Q} f^*_{\text{ext}}(g(X)) \\ &= \sup_{g\in\text{dom}(f^*_{\text{ext}})^{\mathcal{X}},\text{ simple}} \mathbb{E}_{X\sim P} g(X) - \mathbb{E}_{X\sim Q} f^*_{\text{ext}}(g(X)), \end{aligned}$$

where the last step follows since the two suprema combined is equivalent to the supremum over all simple functions g.

□

Remark 2. We remark that when $P \ll Q$ then both results (a) for simple functions and (b) for finite $\mathbb{E}_{X\sim Q} f^*_{\text{ext}}(g(X))$, also hold for supremum over $g : \mathcal{X} \to \mathbb{R}$, i.e. without restricting $g : \mathcal{X} \to \text{dom}(f^*_{\text{ext}})$. As a consequence of the variational characterization, we get the following properties for f-divergences.

1. *Convexity.* First of all, note that $D_f(P\|Q)$ is expressed as a supremum of affine functions since the expectation is a linear operation. As a result, we get that $(P, Q) \mapsto D_f(P\|Q)$ is convex.
2. *Weak lower semicontinuity.* Recall that for an *i.i.d.* zero mean Rademacher vector $X : \Omega \to \{-1, 1\}^m$, the limiting distribution of scaled empirical mean $Y_m \triangleq \frac{1}{\sqrt{m}} \sum_{i=1}^m X_i$ is $\mathcal{N}(0,1)$ as $m \to \infty$ by the central limit theorem. However, $D_f(P_{Y_m}\|\mathcal{N}(0,1)) = f(0) + f'(\infty) > 0$ for all $m \in \mathbb{N}$. This is due to the fact that the former distribution is discrete and the latter is continuous. Therefore similar to the KL divergence, the best we can hope for f-divergence is semicontinuity. Indeed, if $\mathcal{X}$ is a nice space (e.g., Euclidean space), in (1) we can restrict the function g to continuous bounded functions, in which case $D_f(P\|Q)$ is expressed as a supremum of weakly continuous functionals (note that $f^* \circ g$ is also continuous and bounded since f^* is continuous) and is hence weakly lower semicontinuous, i.e., for any sequence of distributions $(P_m \in \mathcal{M}(\mathcal{X}) : m \in \mathbb{N})$ and $(Q_m \in \mathcal{M}(\mathcal{X}) : m \in \mathbb{N})$ such that $P_m \to P$ and $Q_m \to Q$ weakly, we have
$$\liminf_{m\to\infty} D_f(P_m\|Q_m) \geqslant D_f(P\|Q).$$
3. *Relation to DPI.* Variational representations can be thought of as extensions of the DPI. As an exercise, one should try to derive the estimate via both the DPI and (6), for any $A \in \mathcal{F}$
$$|P(A) - Q(A)| \leqslant \sqrt{Q(A)\chi^2(P\|Q)}.$$

Example 1.8 (χ^2-divergence). Recall that χ^2 divergence id f divergence for $f(x) \triangleq (x-1)^2$ for each $x \in \mathbb{R}_+$. We can define its convex extension $f_{\text{ext}}(x) \triangleq (x-1)^2$ for each $x \in \mathbb{R}$. Convex conjugate of f_{ext} is defined for each $y \in \mathbb{R}$ as $f^*_{\text{ext}}(y) = \sup_x xy - (x-1)^2$ which is maximized at $x^* = \frac{y}{2} + 1$, and thus $f^*_{\text{ext}}(y) = y + \frac{y^2}{4}$. We observe that $\text{dom}(f^*_{\text{ext}}) = \mathbb{R}$, and thus substituting this f^*_{ext} in (1), yields

$$\chi^2(P\|Q) = \sup_{h:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{X\sim P} h(X) - \mathbb{E}_{X\sim Q}\left[h(X) + \frac{h^2(X)}{4}\right] \overset{(a)}{=} \sup_{g:\mathcal{X}\to\mathbb{R}} 2\mathbb{E}_{X\sim P} g(X) - \mathbb{E}_{X\sim Q} g^2(X) - 1, \quad (5)$$

where (a) follows from a change of variable $g \triangleq \frac{1}{2}h + 1$. We restrict ourselves to the class of affine function $g^{a,b} : \mathcal{X} \to \mathbb{R}$ defined as $g^{a,b}(x) \triangleq ax + b$ for all $x \in \mathcal{X}$, to write the inequality

$$\sup_{g:\mathcal{X}\to\mathbb{R}} 2\mathbb{E}_{X\sim P} g(X) - \mathbb{E}_{X\sim Q} g^2(X) - 1 \geqslant \sup_{a,b\in\mathbb{R}} 2a\mathbb{E}_{X\sim P} X + 2b - a^2\mathbb{E}_{X\sim Q} X^2 - 2ab\mathbb{E}_{X\sim Q} X - b^2 - 1.$$

The supremum on the right hand side is achieved for $a^* \triangleq \frac{\mathbb{E}_{X\sim P}X - \mathbb{E}_{X\sim Q}X}{\mathrm{Var}_Q X}$ and $b^* \triangleq 1 - a^*\mathbb{E}_{X\sim Q}X$ to write $g^{a^*,b^*}(X) = a^*X + b^* = 1 + a^*(X - \mathbb{E}_{X\sim Q}X)$, and obtain the maximum value

$$\sup_{a,b\in\mathbb{R}} 2\mathbb{E}_{X\sim P}(aX + b) - \mathbb{E}_{X\sim Q}(aX + b)^2 - 1 = \frac{(\mathbb{E}_{X\sim P}X - \mathbb{E}_{X\sim Q}X)^2}{\mathrm{Var}_Q X}. \tag{6}$$

Remark 3. The statistical interpretation of (6) is as follows. If a test statistic $h(X)$ is such that the separation between its expectation under P and Q far exceeds its standard deviation, then this suggests the two hypothesis can be distinguished reliably. The representation (6) will turn out useful in statistical applications for deriving the Hammersley-Chapman-Robbins (HCR) lower bound as well as its Bayesian version, and ultimately the Cramér-Rao and van Trees lower bounds.

2 Varational principles for KL divergence

Definition 2.1. For a random variable $X : \Omega \to \mathcal{X}$ defined on space $(\Omega, \mathcal{F})$, a probability distribution $Q \in \mathcal{M}(\mathcal{X})$, and a measurable map $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$, we define a constant ψ_f, a tilted version of Q as $Q^f \in \mathcal{M}(\mathcal{X})$, and a class of functions $\mathcal{C}_Q$, as

$$\psi_f \triangleq \ln \mathbb{E}_{X\sim Q} e^{f(X)}, \quad dQ^f(x) \triangleq e^{f(x)-\psi_f} dQ(x), \quad \mathcal{C}_Q \triangleq \left\{ f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\} : 0 < e^{\psi_f} < \infty \right\}. \tag{7}$$

We denote the class of all bounded continuous functions as $\mathcal{C}_b$.

Theorem 2.2 (Donsker-Varadhan). *For a random variable $X : \Omega \to \mathcal{X}$ defined on $(\Omega, \mathcal{F})$, distributions $P, Q \in \mathcal{M}(\mathcal{X})$, and measurable map $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$, we have*

$$D(P\|Q) = \sup_{f\in\mathcal{C}_Q} \mathbb{E}_{X\sim P} f(X) - \ln \mathbb{E}_{X\sim Q} e^{f(X)}. \tag{8}$$

In particular, if $D(P\|Q) < \infty$ then $\mathbb{E}_{X\sim P} f(X)$ is well-defined and finite for every $f \in \mathcal{C}_Q$. The identity (8) *holds with $\mathcal{C}_Q$ replaced by the class of all $\mathbb{R}$-valued simple functions. If $\mathcal{X}$ is a normal topological space (e.g., a metric space) with the Borel σ-algebra, then identity* (8) *holds with $\mathcal{C}_Q$ replaced by $\mathcal{C}_b$.*

Proof. We will show upper and lower bounds.

(a) $D \geqslant \sup_{f\in\mathcal{C}_Q}$. We can assume for this part that $D(P\|Q) < \infty$, since otherwise there is nothing to prove. Then fix $f \in \mathcal{C}_Q$ and define a probability measure Q^f, a tilted version of Q as defined in (7). Then, $Q^f \ll Q$. We will apply (2.11) next with reference measure $\mu = Q$. Note that according to (2.10) we always have $\mathrm{Log}\frac{e^{f(x)-\psi_f}}{1} = f(x) - \psi_f$ even when $f(x) = -\infty$. Thus, we get from (2.11)

$$\mathbb{E}_{X\sim P}[f(X)] - \psi_f] = \mathbb{E}_{X\sim P}\mathrm{Log}\frac{dQ^f}{dQ} = D(P\|Q) - D(P\|Q^f) \leqslant D(P\|Q).$$

Note that (2.11) also implies that if $D(P\|Q) < \infty$ and $f \in \mathcal{C}_Q$ the expectation $\mathbb{E}_{X\sim P} f$ is well-defined.

(b) $D \leqslant \sup_f$ *with supremum over all simple functions f.* The idea is to just take $f = \ln\frac{dP}{dQ}$; however to handle all cases we proceed more carefully. First, notice that if $P \not\ll Q$ then for some $E \in \sigma(X)$ with $Q(E) = 0 < P(E)$ and $c \to \infty$ taking $f = c\mathbb{1}_E$ shows that both sides of (8) are infinite. Thus, we assume $P \ll Q$. For any partition of $\mathcal{E} \triangleq (E_1, \dots, E_n)$ such that $\mathcal{X} = \cup_{j=1}^n E_j$, we set $f \triangleq \sum_{j=1}^n \mathbb{1}_{E_j} \ln\frac{P(E_j)}{Q(E_j)}$, to obtain

$$\mathbb{E}_{X\sim P} f(X) - \ln \mathbb{E}_{X\sim Q} e^{f(X)} = D(P_{\mathcal{E}}\|Q_{\mathcal{E}}).$$

By Theorem 1.1, we obtain that supremum over simple functions (and thus over $\mathcal{C}_Q$) is at least as large as $D(P\|Q)$.

(c) $D \leqslant \sup_{f \in \mathcal{C}_b}$ *with supremum over all bounded continuous functions* g. We show that for every simple function g there exists a continuous bounded g_0 such that $\mathbb{E}_{X\sim P} g_0 - \ln \mathbb{E}_{X\sim Q} e^{g_0}$ is arbitrarily close to the same functional evaluated at g. To that end we first show that for any $a \in \mathbb{R}$ and measurable $A \subset \mathcal{X}$ there exists a sequence of continuous bounded $f \triangleq (f_n \in \mathbb{R}^{\mathcal{X}} : n \in \mathbb{N})$ such that

$$\lim_n \mathbb{E}_{X\sim P} f_n(X) = aP(A), \qquad \lim_n \mathbb{E}_{X\sim Q} e^{f_n(X)} = e^a Q(A), \tag{9}$$

hold simultaneously, i.e. $f_n \to a\mathbb{1}_A$ in the sense of approximating both expectations. We only consider the case of $a > 0$ below. Let compact F and open U be such that $F \subset A \subset U$ and $\max(P(U) - P(F), Q(U) - Q(F)) \leqslant \epsilon$. Such F and U exist whenever P and Q are *regular* measures. We notice that finite measures on Polish spaces are regular. Then by Urysohn's lemma there exists a continuous function $f_\epsilon : \mathcal{X} \to [0, a]$ equal to a on F and 0 on U^c. Then we have

$$aP(F) \leqslant \mathbb{E}_{X\sim P} f_\epsilon \leqslant aP(U), \qquad e^a Q(F) \leqslant \mathbb{E}_{X\sim Q} e^{f_\epsilon} \leqslant e^a Q(U).$$

Subtracting $aP(A)$ and $e^a Q(A)$ for each of these inequalities, respectively, we see that taking $\epsilon \to 0$ indeed results in a sequence of functions satisfying (9). Similarly, if we want to approximate a general simple function $g = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$ with A_i disjoint and $|a_i| \leqslant a_{\max} < \infty$, and we fix $\epsilon > 0$ and define functions f_i, ϵ approximating $a_i \mathbb{1}_{A_i}$ as above with sets $F_i \subset A_i \subset U_i$, so that $S \triangleq \cup_{i=1}^n (U_i \setminus F_i)$ satisfies $P(S) \vee Q(S) \leqslant n\epsilon$. We also have

$$|f_{i,\epsilon} - g| \leqslant a_{\max} \sum_{i=1}^n \mathbb{1}_{U_i \setminus F_i} \leqslant n a_{\max} \mathbb{1}_S.$$

We then clearly have $|\mathbb{E}_{X\sim P} \sum_{i=1}^n f_{i,\epsilon} - \mathbb{E}_{X\sim P} g| \leqslant a_{\max} n^2 \epsilon$. On the other hand, we also have

$$\sum_{i=1}^n e^{a_i} Q(F_i) \leqslant \mathbb{E}_{X\sim Q} e^{\sum_{i=1}^n f_{i,\epsilon}} \leqslant \mathbb{E}_{X\sim Q} e^g \mathbb{1}_{S^c} + e^{n a_{\max}} Q(S) \leqslant \mathbb{E}_{X\sim Q} e^g + e^{n a_{\max}} n\epsilon.$$

Hence taking $\epsilon \to 0$ the sum $\sum_{i=1}^n f_{i,\epsilon} \to \sum_{i=1}^n a_i \mathbb{1}_{A_i}$ in the sense of both $\mathbb{E}_{X\sim P}(\cdot)$ and $\mathbb{E}_{X\sim Q} e^{\cdot}$.

□

Corollary 2.3. *For space $\mathcal{X}$, distributions $P, Q \in \mathcal{M}(\mathcal{X})$, and measurable map $f \in \mathcal{C}_Q$, we have $D(P\|Q) \geqslant \mathbb{E}_{X\sim P} f(X) - \psi_f$ with the equality achieved for a unique measure $P = Q^f$ when $D(P\|Q)$ is finite.*

Proof. The inequality follows from Theorem 2.2. We observe that $\mathrm{Log}\frac{dQ^f}{dQ} = f - \psi_f$ even when $f = -\infty$. Therefore, if $D(P\|Q) < \infty$, then

$$\mathbb{E}_{X\sim P}[f(X) - \psi_f] = \mathbb{E}_{X\sim P} \ln \frac{dP}{dQ} - \mathbb{E}_{X\sim P} \ln \frac{dP}{dQ^f} = D(P\|Q) - D(P\|Q^f).$$

It follows that $D(P\|Q) < \infty$ iff $\mathbb{E}_{X\sim P} f(X) < \infty$, and $D(P\|Q) = \mathbb{E}_{X\sim P} f(X) - \psi_f$ iff $D(P\|Q^f) = 0$. □

Proposition 2.4 (Gibbs variational principle). *Let $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ be any measurable function and $Q \in \mathcal{M}(\mathcal{X})$. Then $\psi_f = \sup_{P \in \mathcal{M}(\mathcal{X}) : D(P\|Q) < \infty} \mathbb{E}_{X\sim P} f(X) - D(P\|Q)$. If the left-hand side is finite then the unique maximizer of the right-hand side is $P = Q^f$.*

Proof. Consider $P \in \mathcal{M}(X)$ such that $D(P\|Q) < \infty$, then $P \ll Q$. If $\psi_f = -\infty$ then $\mathbb{E}_{X\sim Q} e^{f(X)} = 0$ which implies that $Q\{f = -\infty\} = 1$. Since $P \ll Q$, we obtain that $P\{f = -\infty\} = 1$, and hence both sides of the above equation are equal to $-\infty$. Next, we consider the case when $\psi_f \in \mathbb{R}$. From Corollary 2.3, we have $\psi_f \geqslant \mathbb{E}_{X\sim P} f(X) - D(P\|Q)$, with equality at $P = Q^f$.

Finally, we consider the case when $\psi_f = \infty$. We define a sequence of bounded functions $f_n \triangleq f \wedge n$ for all $n \in \mathbb{N}$. It follows that $(\psi_{f_n} : n \in \mathbb{N})$ is a non-decreasing sequence of finite numbers with limit $\lim_{n\in\mathbb{N}} \psi_{f_n} = \psi_f = \infty$. Since ψ_{f_n} is finite, there exists a distribution $P_n \in \mathcal{M}(\mathcal{X})$ such that $\mathbb{E}_{P_n} f_n(X) - D(P_n\|Q) = \psi_{f_n}$ for each $n \in \mathbb{N}$. Since $f_n \leqslant f$, we obtain

$$\mathbb{E}_{P_n} f(X) - D(P_n\|Q) \geqslant \psi_{f_n}.$$

The result follows from Fatou's lemma by taking $\liminf$ on both sides. □