Lecture-19: Large sample asymptotics

1 Statistical lower bound from data processing

We give an overview of the classical large-sample theory in the setting of *i.i.d.* observations focusing again on the minimax risk. These results pertain to smooth parametric models in fixed dimensions, with the sole asymptotics being the sample size going to infinity. The main result is that, under suitable conditions, the minimax squared error of estimating θ based on *i.i.d.* sample $X : \Omega \to \mathcal{X}^m$ with common distribution $P_\theta \in \mathcal{P}(\Theta)$ and Fisher information matrix $J_F(\theta)$ satisfies

$$R_m^* \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}[\left\|\hat{\theta} - \theta\right\|^2 \mid \theta] = \frac{1 + o(1)}{m} \sup_{\theta \in \Theta} \operatorname{tr} J_F^{-1}(\theta). \tag{1}$$

This is asymptotic characterization of the minimax risk with sharp constant. In high dimensions, such precise results are difficult and rare. We focus primarily on the quadratic risk and assume that $\Theta \subseteq \mathbb{R}^d$ is an open set. We derive several statistical lower bounds from data processing argument. Specifically, we will take a comparison-of-experiment approach by comparing the actual model with a perturbed model. The performance of a given estimator can be then related to the f-divergence via the data processing inequality and the variational representation. We start by discussing the Hammersley-Chapman-Robbins lower bound which implies the well-known Cramér-Rao lower bound. Because these results are restricted to unbiased estimators, we will also discuss their Bayesian version.

2 Hammersley-Chapman-Robbins (HCR) lower bound

Theorem 2.1 (HCR lower bound). *Consider the statistical decision theory simple setting with $\mathcal{Y} = \Theta = \Theta' \triangleq \mathbb{R}$, and quadratic loss function $L : (\theta, \hat{\theta}) \mapsto (\theta - \hat{\theta})^2$. The quadratic risk at any parameter $\theta \in \Theta$ satisfies*

$$R_\theta(\hat{\theta}) = \mathbb{E}_\theta(\theta - \hat{\theta})^2 \geqslant \operatorname{Var}_\theta(\hat{\theta}) \geqslant \sup_{\theta \neq \theta'} \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\chi^2(P_{\theta'} \| P_\theta)}.$$

Proof. Fix two parameters $\theta' \neq \theta \in \Theta$, and denote their corresponding input distributions as $P_X \triangleq P_\theta$ and $Q_X \triangleq P_{\theta'}$, respectively. The estimator $\hat{\theta}(X, U)$ can be represented by a Markov kernel $P_{\hat{\theta}|X} : \mathcal{X} \to \mathcal{M}(\Theta')$. The estimate $\hat{\theta}(X, U)$ is the output and the corresponding distributions for input distributions P_X and Q_X are $P_{\hat{\theta}}$ and $Q_{\hat{\theta}}$, respectively. Then the data processing inequality for f-divergence and the variational representation of χ^2-divergence from Example A.7 implies that

$$\chi^2(P_X \| Q_X) \geqslant \chi^2(P_{\hat{\theta}} \| Q_{\hat{\theta}}) \geqslant \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\operatorname{Var}_\theta(\hat{\theta})}.$$

□

Corollary 2.2 (Cramér-Rao (CR) lower bound). *Under the regularity conditions for parametric family, on (a) the existence of relative density, (b) the existence of continuous derivative of relative density with respect to parameter θ, and (c) the uniform integrability of the ratio of square of derivative of the density and density, we have for any unbiased estimator $\hat{\theta}$ that satisfies $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for all $\theta \in \Theta \subset \mathbb{R}$,*

$$\operatorname{Var}_\theta(\hat{\theta}) \geqslant \frac{1}{J_F(\theta)}. \tag{2}$$

Proof. From HCR lower bound in Theorem 2.1, we get $R_\theta(\hat{\theta}) = \operatorname{Var}_\theta(\hat{\theta}) \geqslant \sup_{\theta' \neq \theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'} \| P_\theta)}$. The result follows by lower bounding the supremum by the limit of $\theta' \to \theta$, and recalling the asymptotic quadratic expansion of χ^2-divergence in the local neighborhood in terms of the Fisher information. □

Exercise 2.3. Show that for vector $y \in \mathbb{R}^d$ and a positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, we have $\sup_{x\in\mathbb{R}^d:x\neq 0} \frac{\langle x,y\rangle^2}{x^\top \Sigma x} = y^\top \Sigma^{-1} y$, where the maxima is achieved at $x^* = \Sigma^{-1} y$.

Remark 1. We note the following for HCR lower bound and CR lower bound.

- Note that the HCR lower bound is based on the χ^2-divergence. We can write a lower bound version based on Hellinger distance which also implies the CR lower bound.
- Both the HCR and the CR lower bounds extend to the multivariate case as follows. Let $\hat{\theta}$ be an unbiased estimator of $\theta \in \Theta \subseteq \mathbb{R}^d$. Assume that its covariance matrix $\mathrm{Cov}_\theta(\hat{\theta}) \triangleq \mathbb{E}_\theta(\hat{\theta}-\theta)(\hat{\theta}-\theta)^\top$ is positive definite. Fix $a \in \mathbb{R}^d$. Applying HCR lower bound to estimand $T(\theta) \triangleq \langle a,\theta\rangle$ and estimator $\hat{T}(X,U) \triangleq \langle a,\hat{\theta}(X,u)\rangle$, we get

$$\chi^2(P_{\theta'}\|P_\theta) \geqslant \frac{(\mathbb{E}_\theta\langle a,\hat{\theta}\rangle - \mathbb{E}_{\theta'}\langle a,\hat{\theta}\rangle)^2}{\mathrm{Var}_\theta\langle a,\hat{\theta}\rangle} = \frac{\langle a,\theta-\theta'\rangle^2}{a^\top \mathrm{Cov}_\theta(\hat{\theta})a}.$$

  Since the choice of $a \in \mathbb{R}^d$ was arbitrary, the right hand side of the equation holds for all a. Taking supremum over a, it follows from Exercise 2.3 that

$$\chi^2(P_{\theta'}\|P_\theta) \geqslant (\theta-\theta')^\top \mathrm{Cov}_\theta(\hat{\theta})^{-1}(\theta-\theta').$$

- From the additivity property of the Fisher information, the Fisher information matrix for a sample of m *i.i.d.* observations is equal to $mJ_F(\theta)$. Writing the Taylor series expansion of χ^2-divergence in the neighborhood of $\theta \in \Theta \subseteq \mathbb{R}^d$, we get

$$(\theta'-\theta)^\top \Big(mJ_F(\theta) - (\mathrm{Cov}_\theta(\hat{\theta}))^{-1}\Big)(\theta'-\theta) + o(\|\theta'-\theta\|^2) \geqslant 0.$$

  Taking the limit $\theta' \to \theta$, we obtain $mJ_F(\theta) - (\mathrm{Cov}_\theta(\hat{\theta}))^{-1} \succeq 0$, and taking trace we conclude that the squared error of any unbiased estimators satisfies

$$\mathbb{E}_\theta\|\hat{\theta}-\theta\|^2 = \mathrm{tr}\,\mathrm{Cov}_\theta(\hat{\theta}) \geqslant \frac{1}{m}\,\mathrm{tr}\,J_F^{-1}(\theta).$$

  This is already very close to (1), except for the fundamental restriction of unbiased estimators.

A Variational representation of f-divergences

Theorem A.1. *Let $P,Q \in \mathcal{M}(\mathcal{X})$. Given finite partition $\mathcal{E} \triangleq \{E_1,\ldots,E_n\} \subseteq \mathcal{F}$ 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*

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

Definition A.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 for all $y\in\mathbb{R}$ as

$$f^*(y) \triangleq \sup_{x\in\mathbb{R}_+} xy - f(x).$$

The domain of convex conjugate f^* is denoted by $\mathrm{dom}(f^*) \triangleq \{y\in\mathbb{R} : f^*(y) < \infty\}$.

Lemma A.3. *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 semi-continuous.*

Proof. Recall that f^* is the convex conjugate of f.
(a) Since the supremum of affine maps is convex, it follows that f^* is convex.
(b) Since $zy - f^*(y)$ is concave, it has a unique maximum. From definition of convex conjugate, we have $f(x) \geqslant xy - f^*(y)$ for all $y\in\mathbb{R}$, hence $f^{**}(x) \leqslant f(x)$ for all $x\in\mathbb{R}_+$.

□

Definition A.4. 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

$$\Psi^*(g) \triangleq \sup_P \mathbb{E}_P g(X) - \Psi(P).$$

Remark 2. Under appropriate conditions e.g. finite $\mathcal{X}$, biconjugation yields the sought-after variational representation

$$\Psi(P) = \sup_g \mathbb{E}_P g(X) - \Psi^*(g).$$

Next 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 A.5. 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 3. 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 A.6. *Let $P, Q \in \mathcal{M}(\mathcal{X})$. Consider a convex function $f : (0,\infty) \to \mathbb{R}$, its extension f_{ext} and its convex conjugate f^*_{ext}. Then,*

$$D_f(P\|Q) = \sup_{g:\mathcal{X}\to \text{dom}(f^*_{\text{ext}})} \mathbb{E}_P g(X) - \mathbb{E}_Q f^*_{\text{ext}}(g(X)), \tag{3}$$

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

Proof. We will show this in three steps.

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

$$\mathbb{E}_P g(X) \leqslant D_f(P\|Q) + \mathbb{E}_Q f^*_{\text{ext}}(g(X)). \tag{4}$$

We denote the densities of P and Q by p, q respectively. Then, from the definition of f^*_{ext} we have for every $x \in \{z \in \mathcal{X} : q(z) > 0\}$,

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

Integrating this over $dQ = q d\mu$ restricted to the set $\{q > 0\}$, we get

$$\mathbb{E}_Q f^*_{\text{ext}}(g(X)) + \int_{x\in\mathcal{X}:q(x)>0} d\mu(x) q(x) f_{\text{ext}}\Big(\frac{p(x)}{q(x)}\Big) \geqslant \mathbb{E}_P g(X) \mathbb{1}_{\{q(X)>0\}}. \tag{5}$$

We notice that $\sup\{y \in \mathbb{R} : y \in \text{dom}(f^*_{\text{ext}})\} = \lim_{x\to\infty} \frac{f_{\text{ext}}(x)}{x} = f'(\infty)$. Therefore, $f'(\infty) P\{q(X) = 0\} \geqslant \mathbb{E}_P g(X) \mathbb{1}_{\{q(X)=0\}}$. Summing this inequality with (5) we obtain the desired result in (4).

2. **Step 2.** We prove that supremum in (3) over simple functions g does yield $D_f(P\|Q)$, so that inequality (4) is tight. From Theorem A.1, it suffices to show (3) for finite observation space $\mathcal{X}$. Indeed, for general $\mathcal{X}$, given a finite partition $\mathcal{E} \triangleq \{E_1, \ldots, E_n\}$ 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

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

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

3. **Step 3.** We consider finite $\mathcal{X}$. Let $S \triangleq \{x \in \mathcal{X} : Q(x) > 0\}$ denote the support of Q. We show the following statement which is equivalent to (3),

$$D_f(P\|Q) = \sup_{g:S\to \mathrm{dom}(f^*_{\mathrm{ext}})} \mathbb{E}_P g(X) - \mathbb{E}_Q[f^*_{\mathrm{ext}}(g(X))] + f'(\infty)P(S^c). \tag{6}$$

Defining functional $\Psi(P) \triangleq \sum_{x\in S} Q(x) f_{\mathrm{ext}}\left(\frac{P(x)}{Q(x)}\right)$ where P takes values over all signed measures on S, we have

$$D_f(P\|Q) = \Psi(P) + f'(\infty)P(S^c).$$

Functional $\Psi(P)$ can be identified with $\mathbb{R}^S$. The convex conjugate of $\Psi(P)$ is defined for any $g : S \to \mathbb{R}$, as

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

where (a) follows from the minimax theorem which applies due to finiteness of $\mathcal{X}$. It follows that

$$\Psi^*(g) = \mathbb{E}_Q f^*_{\mathrm{ext}}(g(X))\mathbb{1}_{\left\{g\in \mathrm{dom}(f^*_{\mathrm{ext}})^S\right\}} + \infty \mathbb{1}_{\left\{g\notin \mathrm{dom}(f^*_{\mathrm{ext}})^S\right\}}$$

Recall that if Ψ is convex, then the convex conjugate Ψ^* is Ψ itself. Applying the convex duality of convex conjugates yields the proof of the desired (6).

□

Remark 4. We remark that when $P \ll Q$ then both results (a) and (b) also hold for supremum over $g : \mathcal{X} \to \mathbb{R}$, i.e. without restricting $g(x) \in \mathrm{dom}(f^*_{\mathrm{ext}})$. 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 $Y_m \triangleq \frac{1}{\sqrt{m}}\sum_{i=1}^m X_i$ 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 (3) 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

$$\liminf_{m\to\infty} D_f(P_n\|Q_m) \geqslant D_f(P\|Q).$$

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 (8), for any $A \in \mathcal{F}$

$$|P(A) - Q(A)| \leqslant \sqrt{Q(A)\chi^2(P\|Q)}.$$

Example A.7 (χ^2-divergence). For χ^2-divergence we have $f(x) = (x-1)^2$. Take $f_{\mathrm{ext}}(x) = (x-1)^2$, whose conjugate is $f^*_{\mathrm{ext}}(y) = \sup_{x\in\mathbb{R}_+} xy - (x-1)^2$ which is maximized at $x = \frac{y}{2} + 1$, and for which $f^*_{\mathrm{ext}}(y) = y + \frac{y^2}{4}$. Applying (3) yields

$$\chi^2(P\|Q) = \sup_{h:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P h(X) - \mathbb{E}_Q\left[h(X) + \frac{h^2(X)}{4}\right] = \sup_{g:\mathcal{X}\to\mathbb{R}} 2\mathbb{E}_P g(X) - \mathbb{E}_Q g^2(X) - 1, \tag{7}$$

where the last step 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

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

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

$$\sup_{a,b\in\mathbb{R}} 2\mathbb{E}_P(aX+b) - \mathbb{E}_Q(aX+b)^2 - 1 = 2a^*(\mathbb{E}_P X - \mathbb{E}_Q X) - (a^*)^2 \mathrm{Var}_Q X = \frac{(\mathbb{E}_P X - \mathbb{E}_Q X)^2}{\mathrm{Var}_Q X}. \quad (8)$$

Remark 5. The statistical interpretation of (8) 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 (8) 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.

B Varational principles for KL divegrence

Definition B.1. For space $\mathcal{X}$, 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 for all $x \in \mathcal{X}$, and a class of functions $\mathcal{C}_Q$, as

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

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

Theorem B.2 (Donsker-Varadhan). *For space $\mathcal{X}$, distributions $P, Q \in \mathcal{M}(\mathcal{X})$, and measurable map $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$, we have*

$$D(P\|Q) = \sup_{f\in\mathcal{C}_Q} \mathbb{E}_P f(X) - \ln \mathbb{E}_Q e^{f(X)}. \quad (9)$$

In particular, if $D(P\|Q) < \infty$ then $\mathbb{E}_P f(X)$ is well-defined and finite for every $f \in \mathcal{C}_Q$. The identity (9) *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* (9) *holds with $\mathcal{C}_Q$ replaced by $\mathcal{C}_b$.*

Corollary B.3. *For space $\mathcal{X}$, distributions $P, Q \in \mathcal{M}(\mathcal{X})$, and measurable map $f \in \mathcal{C}_Q$, we have $D(P\|Q) \geqslant \mathbb{E}_P f(X) - \psi_f$ with the equality achieved for a unique measure $P = Q^f$ when $D(P\|Q)$ is finite.*

Proof. The inequality follows from Theorem B.2. We observe that $\ln \frac{dQ^f}{dQ} = f - \psi_f$. Therefore,

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

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

Proposition B.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 $\psi_f = \sup_{P\in\mathcal{M}(\mathcal{X}):D(P\|Q)<\infty} \mathbb{E}_P f(X) - D(P\|Q)$. 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 $\mathbb{E}_Q e^{f(X)} = 0$ which implies that $Q\{f = -\infty\} = 1$, and hence $P\{Q = -\infty\} = 1$. In turn both sides of the above equation are equal to $-\infty$. Next, we consider the case when $\psi_f \in \mathbb{R}$. From Corollary B.3, we have $\psi_f \geqslant \mathbb{E}_P f(X) - D(P\|Q)$, 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 lim inf on both sides. □