DATA 37200 HW1, Winter 2026

Problem 1

We go through the proofs of some useful properties of the KL divergence. For simplicity, consider discrete distributions P and Q in what follows.

(a) Show that for any function f,

$$D_{KL}(\mathcal{L}_P(f(X)), \mathcal{L}_Q(f(X))) \leq D_{KL}(P, Q).$$

This is called the "data processing inequality". Here $\mathcal{L}_P(f(X))$ is notation for the law/distribution of $f(X)$ when $X \sim P$. (Hint: it can be done by applying Jensen's inequality.)

(b) A version of Bretagnolle-Huber inequality: for any event A

$$P(A) + Q(A^C) \geq 1 - \sqrt{1 - \exp(-D_{KL}(P, Q))}.$$

For this part, please feel free to refer to chapter 14 of the Lattimore-Szepesvari textbook which explains the steps. Just write a self-contained proof of the inequality stated above.

(c) Compute the KL divergence $D_{KL}(P, Q)$ between $P = N(\mu_1, \sigma_1^2)$ and $Q = N(\mu_2, \sigma_2^2)$.

Problem 2

The Donsker-Varadhan variational formula characterizes the KL divergence as the convex conjugate ("Legendre transform") of the cumulant generating function. Let P and Q be probability measures on a discrete set $\mathcal{X}$. We wish to prove:

$$D_{KL}(P \| Q) = \sup_f \left\{ \mathbb{E}_P[f(X)] - \ln \mathbb{E}_Q[e^{f(X)}] \right\}.$$

(a) The Lower Bound: Let f be any function. Define a new probability measure Q_f with the Radon-Nikodym derivative $\frac{dQ_f}{dQ} = \frac{e^f}{\mathbb{E}_Q[e^f]}$. By considering the non-negativity of $D_{KL}(P\|Q_f)$, show that:

$$D_{KL}(P\|Q) \geq \mathbb{E}_P[f] - \ln \mathbb{E}_Q[e^f]$$

(b) The Optimal Function: Assume $P \ll Q$ and let $g = \frac{dP}{dQ}$ be the density. Define $f^* = \ln g$. Show that f^* achieves the value $D_{KL}(P\|Q)$.

Problem 3

Pinsker's inequality is a fundamental result in information theory that bounds the Total Variation distance between two probability measures by their Kullback-Leibler divergence:

$$TV(P,Q) \leq \sqrt{\frac{1}{2} D_{KL}(P\|Q)}$$

Complete the following steps to derive this inequality using the properties of the Bernoulli distribution.

(a) Let $X \sim \text{Ber}(p)$ for $p \in (0,1)$. Prove that the cumulant generating function (CGF) $\psi_p(\lambda) = \ln \mathbb{E}[e^{\lambda X}]$ is:

$$\psi_p(\lambda) = \ln(1 - p + pe^{\lambda})$$

(b) Define $\phi_p(\lambda) = \psi_p(\lambda) - \lambda p$ as the CGF of the centered Bernoulli variable. Use a Taylor expansion of $\phi_p(\lambda)$ around $\lambda = 0$ to show that:

$$\phi_p(\lambda) \leq \frac{\lambda^2}{8}$$

(Note: we have already done this in class, but please give a self-contained proof here for practice. Feel free to refer to notes/textbook.)

(c) By problem 2, the KL divergence between two Bernoulli distributions is the "Legendre transform" of the CGF: $D_{KL}(p\|q) = \sup_{\lambda \in \mathbb{R}} \{\lambda p - \psi_q(\lambda)\}$. Use the quadratic bound from part (b) to prove:

$$D_{KL}(p\|q) \geq 2(p-q)^2$$

(d) Use the Data Processing Inequality for KL divergence and the definition $TV(P, Q) = \sup_A |P(A) - Q(A)|$ to extend the Bernoulli result to any two probability measures P and Q:

$$D_{KL}(P, Q) \geq 2TV(P, Q)^2.$$