Lecture 7: Data Processing Inequality

We will introduce the data processing inequality (DPI) here. In essence, the DPI shows that the $f$-divergence between two distributions does not decrease when we push it through a transition kernel. This can be thought of as follows: pushing two observations $X$ and $Y$ through a channel will only make it harder to distinguish between them.

$$
\begin{align*}
P_X & \quad \xrightarrow{P_{Y|X}} \quad P_Y \\
Q_X & \quad \xrightarrow{P_{Y|X}} \quad Q_Y
\end{align*}
$$

**Theorem 1 (Data Processing Inequality)** Let $P_X, Q_X \in \mathcal{P}(\mathcal{X})$ and $P_{Y|X}$ be a transition kernel from $(\mathcal{X}, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{G})$. Let $P_Y, Q_Y \in \mathcal{P}(\mathcal{Y})$ be the transformation of $P_X$ and $Q_X$, respectively, when pushed through $P_{Y|X}$, i.e., $P_X(B) = \int_X P_{Y|X}(B \mid x) dP_X(x)$. Then, for any $f$-divergence, we have that

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

**Example 7.1**

1. **Gaussian Convolutions:** Let $X \sim P_X$, $X' \sim Q_X$, and $Z, Z' \sim N_\sigma := \mathcal{N}(0, \sigma^2 I_d)$ be independent random variables. Define $Y := X + Z$ and $Y' := X' + Z'$. Here, the transition kernel is $P_{Y|X}(\cdot \mid x) = \mathcal{N}(x, \sigma^2 I_d)$. Recall that for two independent random variables $W \sim \mu$ and $W' \sim \nu$, it holds that $W + W' \sim \mu * \nu$ where $\mu * \nu$ is the convolution of $\mu$ and $\nu$ defined as

$$(\mu * \nu)(A) = \int_A \int_Y 1_{\{x+y\in A\}} d\mu(x) d\nu(y),$$

for any measurable $A \subseteq \mathcal{X} + \mathcal{Y} := \{x + y : x \in \mathcal{X}, y \in \mathcal{Y}\}$ (note that if $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, then $\mu * \nu \in \mathcal{P}(\mathcal{X} + \mathcal{Y})$).

It follows that $Y \sim P_X * N_\sigma$ and $Y' \sim Q_X * N_\sigma$. The DPI implies

$$D_f(P_X \| Q_X) \geq D_f(P_X * N_\sigma \| Q_X * N_\sigma).$$

2. **Deterministic Functions:** Let $X \sim P_X$, $X' \sim Q_X$ and set $Y = g(X)$, $Y' = g(X')$ for a deterministic measurable $g$. The transition kernel is $P_{Y|X}(\cdot \mid x) = \delta_{g(x)}(\cdot)$, where $\delta_a$ is the Dirac measure centered at $a \in \mathcal{X}$, i.e., $\delta_a(A) = 1_A(a)$, for any $A$ measurable.

   (i) Let $E$ be any measurable event, define $g(x) = 1_{\{x \in E\}}$ and set $Y = g(X)$. Note that $Y$ is a binary random variable with $P_Y(\{1\}) = P_X(E)$. This implies that $Y = 1_{\{X \in E\}} \sim \text{Ber}(P_X(E))$ and $Y' = 1_{\{X' \in E\}} \sim \text{Ber}(Q_X(E))$. By the data processing inequality, we obtain

$$D_f(P_X \| Q_X) \geq D_f\left(\text{Ber}(P_X(E)) \| \text{Ber}(Q_X(E))\right),$$

for all measurable $E$. 

(ii) Consider \( g(x_1, x_2) = x_1 \), and let \( X = (X_1, X_2) \sim P_{X_1,X_2}, X' = (X'_1, X'_2) \sim Q_{X_1,X_2} \), \( Y = g(X_1, X_2) = X_1 \), and \( Y' = g(X'_1, X'_2) = X'_1 \). It follows that \( Y \sim P_{X_1} \) and \( Y \sim Q_{X_1} \). Applying the data processing inequality gives

\[
D_f(P_{X_1,X_2} \| Q_{X_1,X_2}) \geq D_f(P_{X_1} \| Q_{X_1})
\]

By Item (iv) from the properties of \( f \)-divergences we have that if \( P_{X_2|X_1} = Q_{X_2|X_1} \) then equality above holds.

**Proof of DPI:** Throughout this proof we use the shorthand \( \frac{dP}{dQ} = \frac{dP}{d\lambda} \frac{d\lambda}{dQ} \), where \( \lambda \) is a measure that dominates both \( P \) and \( Q \) (e.g., \( \lambda = P + Q \)), and \( dP/d\lambda \) is the Radon-Nikodym derivative of \( P \) w.r.t. \( \lambda \).

First, recall that if \( P_{XY} = P_X P_{Y|X} \) and \( Q_{XY} = Q_X P_{Y|X} \), then

\[
D_f(P_X \| Q_X) = D_f(P_{XY} \| Q_{XY}) = \mathbb{E}_{Q_{XY}} \left[ f \left( \frac{dP_{XY}}{dQ_{XY}} \right) \right].
\]

Using the law of total expectation, we get

\[
\mathbb{E}_{Q_{XY}} \left[ f \left( \frac{dP_{XY}}{dQ_{XY}} \right) \right] = \mathbb{E}_{Q_Y} \left[ \mathbb{E}_{Q_{X|Y}} \left[ f \left( \frac{dP_{XY}}{dQ_{XY}} \right) \bigg| Y \right] \right].
\]

As \( f \) is convex, applying Jensen’s inequality yields

\[
\mathbb{E}_{Q_Y} \left[ \mathbb{E}_{Q_{X|Y}} \left[ f \left( \frac{dP_{XY}}{dQ_{XY}} \right) \bigg| Y \right] \right] \geq \mathbb{E}_{Q_Y} \left[ f \left( \mathbb{E}_{Q_{X|Y}} \left[ \frac{dP_{XY}}{dQ_{XY}} \bigg| Y \right] \right) \right].
\]

To conclude the proof, it suffices to show that

\[
\mathbb{E}_{Q_{X|Y}} \left[ \frac{dP_{XY}}{dQ_{XY}} \bigg| Y \right] = \frac{dP_Y}{dQ_Y}.
\]

It holds that

\[
\mathbb{E}_{Q_{X|Y}} \left[ \frac{dP_{XY}}{dQ_{XY}} \bigg| Y \right] = \int_X \frac{dP_{XY}}{dQ_{XY}} dQ_X|Y = \int_X \frac{dP_Y dP_{X|Y}}{dQ_Y dQ_X|Y} dQ_X|Y = \int_X \frac{dP_Y}{dQ_Y} dP_{X|Y} = \frac{dP_Y}{dQ_Y}.
\]

■