1. **Entropy (full) Chain Rule.**
   Let $(X_1, ..., X_k) \sim P_{X_1,...,X_k}$. Show that:

   (a) If $(X_1, ..., X_k)$ is discrete, then its Shannon entropy decomposes as

$$H(X_1, ..., X_k) = \sum_{i=1}^{k} H(X_i \mid X_1, ..., X_{i-1})$$

   where $H(X_1 \mid X_0) = H(X_1)$.

   Solution.

$$H(X_1, ..., X_k) := \mathbb{E}_P \left[\log \left(\frac{1}{P_{X_1,...,X_k}(X_1, ..., X_k)} \right) \right]$$

$$\begin{aligned}
&= \mathbb{E}_P \left[\log \left(\frac{1}{P_{X_1} P_{X_2|X_1} P_{X_3|X_2,X_1} \dots P_{X_i|X_1,...,X_{i-1}} \dots P_{X_k|X_1,...,X_{k-1}}} \right) \right] \\
&= \mathbb{E}_P \left[\log \left(\frac{1}{P_{X_1}} \right) + \log \left(\frac{1}{P_{X_2|X_1}} \right) + ... + \log \left(\frac{1}{P_{X_k|X_1,...,X_k}} \right) \right] \\
&= H(X_1 \mid X_0) + H(X_2 \mid X_1, X_2) + ... + H(X_k \mid X_1, ..., X_{k-1}) \\
&= \sum_{i=1}^{k} H(X_i \mid X_1, ..., X_{i-1})
\end{aligned}$$

■

   (b) If $(X_1, ..., X_k)$ is jointly continuous, then its differential entropy decomposes as

$$h(X_1, ..., X_k) = h(X_k) + \sum_{i=1}^{k-1} h(X_{k-i} \mid X_k, ..., X_{k-i+1}).$$

Solution.

$$
\begin{aligned}
h(X_1, ..., X_k) &:= \mathbb{E}_P\left[\log\left(\frac{1}{P_{X_1,...,X_k}(X_1,...,X_k)}\right)\right] \\
&= \mathbb{E}_P\left[\log\left(\frac{1}{P_{X_k}P_{X_{k-1}|X_k}P_{X_{k-2}|X_k,X_{k-1}}\dots P_{X_{k-i}|X_k,...,X_{k-i+1}}\dots P_{X_1|X_k,...,X_2}}\right)\right] \\
&= \mathbb{E}_P\left[\log\left(\frac{1}{P_{X_k}}\right) + \log\left(\frac{1}{P_{X_{k-1}|X_k}}\right) + ... + \log\left(\frac{1}{P_{X_1|X_k,...,X_2}}\right)\right] \\
&= h(X_k) + h(X_{k-1} \mid X_k) + ... + H(X_1 \mid X_k, ..., X_2) \\
&= \sum_{i=1}^{k} H(X_i \mid X_1, ..., X_{i-1})
\end{aligned}
$$

■

2. **Properties of Mutual Information.**
Let $(X, Y, Z) \sim P_{X,Y,Z} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$. Establish the following relations:

   (a) KL Divergence Chain Rule: For any $Q_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, we have

$$
D_{\mathsf{KL}}\left(P_{X,Y} \| Q_{X,Y}\right) = D_{\mathsf{KL}}\left(P_X \| Q_X\right) + D_{\mathsf{KL}}\left(P_{Y|X} \| Q_{Y|X} \mid P_X\right)
$$

Solution.

$$
\begin{aligned}
D_{\mathsf{KL}}\left(P_{X,Y} \| Q_{X,Y}\right) &= \mathbb{E}_{Q_{XY}}\left[\frac{\mathsf{d}P_{XY}}{\mathsf{d}Q_{XY}} \log \frac{\mathsf{d}P_{XY}}{\mathsf{d}Q_{XY}}\right] \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \frac{\mathsf{d}P_{XY}}{\mathsf{d}Q_{XY}} \log \frac{\mathsf{d}P_{XY}}{\mathsf{d}Q_{XY}} \mathsf{d}Q_{XY} \\
&= \mathsf{d}P_{XY} \log \frac{\mathsf{d}P_{XY}}{\mathsf{d}Q_{XY}} \\
&= \mathsf{d}P_{Y|X}\mathsf{d}P_X \log \frac{\mathsf{d}P_{Y|X}\mathsf{d}P_X}{\mathsf{d}Q_{Y|X}\mathsf{d}Q_X} \\
&= \mathsf{d}P_{Y|X}\mathsf{d}P_X \log\left(\frac{P_X}{Q_X}\right) - \mathsf{d}P_{Y|X}\mathsf{d}P_X \log\left(\frac{\mathsf{d}P_{Y|X}}{\mathsf{d}Q_{Y|X}}\right) \\
&= D_{\mathsf{KL}}\left(P_X \| Q_X\right) + D_{\mathsf{KL}}\left(P_{Y|X} \| Q_{Y|X} \mid P_X\right)
\end{aligned}
$$

■

(b) Relation to Conditional KL Divergence: $I(X;Y) = D_{\mathsf{KL}}\left(P_{Y|X} \| Q_{Y|X} \mid P_X\right)$, where $P_{X,Y} = P_X P_{Y|X}$ and P_Y is the Y-marginal.

Solution.

$$\begin{aligned} I(X;Y) &= D_{\mathsf{KL}}\left(P_{XY} \| P_X \times P_Y\right) \\ &= D_{\mathsf{KL}}\left(P_X P_{Y|X} \| P_X \times P_Y\right) \\ &= D_{\mathsf{KL}}\left(P_X \| P_X\right) + D_{\mathsf{KL}}\left(P_{Y|X} \| P_Y \mid P_X\right) \\ &= D_{\mathsf{KL}}\left(P_{Y|X} \| P_Y \mid P_X\right) \end{aligned}$$

■

(c) Symmetry: $I(X;Y) = I(Y;X)$.

Solution. We apply the data processing inequality in both directions. Let $f(x,y) = (y,x)$ be our transition kernel. Then

$$\begin{aligned} D_{\mathsf{KL}}\left(P_{XY} \| P_X \times P_Y\right) &\le D_{\mathsf{KL}}\left(P_{YX} \| P_Y \times P_X\right) \\ D_{\mathsf{KL}}\left(P_{YX} \| P_Y \times P_X\right) &\le D_{\mathsf{KL}}\left(P_{XY} \| P_X \times P_Y\right) \\ \rightarrow D_{\mathsf{KL}}\left(P_{XY} \| P_X \times P_Y\right) &= D_{\mathsf{KL}}\left(P_{YX} \| P_Y \times P_X\right) \end{aligned}$$

■

(d) More Data $\implies$ More Information: $I(X;Y) \le I(X;Y,Z)$.

Solution. First we expand $I(X;Y,Z)$.

$$\begin{aligned} I(X;Y,Z) &= D_{\mathsf{KL}}\left(P_{XYZ} \| P_X \times P_{YZ}\right) \\ &= D_{\mathsf{KL}}\left(P_{YZ|X} \| P_{YZ} \mid P_X\right) \\ &= \int P_X P_{Y|X} P_{Z|XY} \log\left(\frac{P_{Y|X} P_{Z|XY}}{P_Y P_{Z|Y}}\right) \\ &= \int P_X P_{Y|X} P_{Z|XY} \left[\log\left(\frac{P_{Y|X}}{P_Y}\right) + \log\left(\frac{P_{Z|XY}}{P_{Z|Y}}\right)\right] \\ &= \int P_X P_{Y|X} P_{Z|XY} \log\frac{P_{Y|X}}{P_Y} + P_X P_{Y|X} P_{Z|XY} \log\frac{P_{Z|XY}}{P_{Z|Y}} \\ &= D_{\mathsf{KL}}\left(P_{Y|X} \| P_Y \mid P_X\right) + D_{\mathsf{KL}}\left(P_{Z|XY} \| P_{Z|Y} \mid P_{XY}\right) \\ &\ge D_{\mathsf{KL}}\left(P_{Y|X} \| P_Y \mid P_X\right) \\ &= I(X;Y) \end{aligned}$$

Note: Letting f(x,y,z) = (x,y) and using data processing inequality also suffices for the proof and avoids the mess that is above. ■

(e) Mutual Information and Functions: $I(X;Y) \geq I(X;f(Y))$ for any deterministic function f. Furthermore, if f is continuous and one-to-one, then

$$I(X;f(X)) = \begin{cases} H(X), & \text{if X is discrete} \\ \infty, & \text{if X is continuous} \end{cases}.$$

Solution.

$$I(X;Y) = D_{\mathsf{KL}}\left(P_{XY} \| P_X \times P_Y\right)$$

Using DPI for the kl divergence, with the transition kernel that maps Y to $f(Y)$, we have $D_{\mathsf{KL}}\left(P_{XY} \| P_X \times P_Y\right) \geq D_{\mathsf{KL}}\left(P_{Xf(Y)} \| P_X \times P_{f(Y)}\right) = I(X;f(Y))$ which gives us the desired result of $I(X;Y) \geq I(X;f(Y))$
Furthermore, if f is continuous and one-to-one then for a discrete X we have

$$\begin{aligned} I(X;f(X)) &= D_{\mathsf{KL}}\left(P_{X|f(X)} \| P_X \mid P_{f(X)}\right) \\ &= \sum_{x\in\mathcal{X}} p_X(x) \sum_{x'\in\mathcal{X}} \delta_X(x') \frac{\log(\delta_X(x'))}{p_X(x')} \\ &= \sum_{x\in\mathcal{X}} p_X(x) \log \frac{1}{p_X(x)} = H(X). \end{aligned}$$

If X is instead continuous we show that $P_{XfX} \not\ll\not\ll P_X \times P_{f(X)}$. Let $\nabla = \{(x, f(x)) \mid x \in \mathcal{X}\}$.

$$P_{Xf(X)}(\nabla) = \int_{\nabla} \mathsf{d}P_{Xf(X)} = \int_{x\in\mathcal{X}} \mathsf{d}P_X(x)\mathsf{d}\int_{x'\in\mathcal{X}} \mathsf{d}\delta_X(x') = 1 > 0$$

However

$$P_X \times P_{f(X)}(\nabla) = \int_{\nabla} \mathsf{d}P_X \times P_{f(X)}(s, x') = 0$$

which will cause the kl divergence to blow up to infinity. ■

3. **Entropy of a Sum.**
Let $(X,Y) \sim P_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, where $\mathcal{X} = \{x_1, ..., x_r\}$ and $\mathcal{Y} = \{y_1, ..., y_s\}$, and define $Z = X + Y$.

(a) Show that $\max\{H(X), H(Y)\} \leq H(Z) \leq H(X) + H(Y)$ when X is independent of Y.

Solution.

$$\begin{aligned} H(Z) = H(X+Y) &\leq H(X+Y, Y) && \text{follows from the chain rule} \\ &= H(X) + H(X+Y \mid X) \\ &= H(X) + H(Y \mid X) && \text{see part b} \\ &\leq H(X) + H(Y) \end{aligned}$$

If X and Y are independent, we have

$$\begin{aligned} H(Z) &\geq H(Z|X) && && (1)\\ &= H(Y|X) && \text{see part b} && (2)\\ &= H(Y) && \text{since } Y \text{ and } X \text{ are independent} && (3) \end{aligned}$$

Similarly, we have $H(Z) \geq H(X)$, which gives us $\max(H(Y), H((X)) \leq H(Z)$ ■

(b) Show that $H(Z \mid X) = H(Y \mid X)$. Argue that if (X, Y) are independent, then $H(Y) \leq H(Z)$ and $H(X) \leq H(Z)$. Thus the summation of *independent* random variables increases uncertainty.

Solution.

$$\begin{aligned} H(Z \mid X) &:= \mathbb{E}_{XZ}\left[\log\left(\frac{1}{P_{Z|X}}\right)\right] \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathbb{P}(Z = x+y, X = x)\log\frac{1}{\mathbb{P}(Z = x+y|X = x)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathbb{P}(Y = y, X = x)\log\frac{1}{\mathbb{P}(Y = y|X = x)} \\ &= H(Y|X) \end{aligned}$$

■

(c) Give an example of dependent random variables for which $H(X) > H(Z)$ and $H(Y) > H(Z)$.

Solution. Consider $X = Ber(\frac{1}{2})$, and $Y = -X$. That will give us $Z = 0$ (the constant RV). We get

$$1 = H(X) = H(Y) > H(Z) = 0$$

■

4. **Information Inequalities.**
Let $(X, Y, Z) \sim P_{X,Y,Z} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$. Prove the following inequalities and find (necessary and sufficient) conditions for equality.

(a) $H(X, Y \mid Z) \geq H(X \mid Z)$.

Solution.

$$\begin{aligned}
H(X, Y \mid Z) &= \mathbb{E}\left[\log \frac{1}{P_{XY|Z}}\right] \\
&= \mathbb{E}\left[\log \frac{1}{P_{X|Z} P_{Y|XZ}}\right] \\
&= \mathbb{E}\left[\log \frac{1}{P_{X|Z}}\right] + \mathbb{E}\left[\log \frac{1}{P_{Y|XZ}}\right] \\
&= H(X \mid Z) + H(Y \mid X, Z) \geq H(X \mid Z).
\end{aligned}$$

From the above we see that equality holds when Y is completely determined once given X, Z. ∎

(b) $I(X, Y; Z) \geq I(X; Z)$.

Solution. This directly follows from 2d – more data leads to more information. The equality condition is if $H(X|Y, Z) = H(X|Z)$. In other words, given Z, Y does not provide more information about X. ∎

(c) $H(X, Y, Z) - H(X, Y) \leq H(X, Z) - H(X)$.

Solution. Use the decompositions

$$\begin{gathered}
H(X, Y, Z) = H(X, Y) + H(Z \mid X, Y) \to H(X, Y, Z) - H(X, Y) = H(Z \mid X, Y) \\
H(X, Z) = H(X) + H(Z \mid X) \to H(X, Z) - H(X) = H(Z \mid X)
\end{gathered}$$

We are left to prove $H(Z \mid X, Y) \leq H(Z \mid X)$. This follows from the property that conditioning decreases entropy.
The condition for equality thus becomes Z being conditionally independent of Y given X. ∎

(d) $I(X; Z \mid Y) \geq I(Z; Y \mid X) - I(Z; Y) + I(X; Z)$.

Solution. We will start by expanding both sides.

$$LHS = H(Z|Y) - H(Z|X, Y)$$

and

$$\begin{aligned}
RHS &= H(Z|X) - H(Z|X, Y) - H(Z) + H(Z|Y) + H(Z) - H(Z|X) \\
&= H(Z|Y) - H(Z|X, Y)
\end{aligned}$$

So we have LHS = RHS, and the inequality holds as a strict equality *always* ∎

5. **Shannon Entropy on Infinite Alphabets.**
Let $X \sim P \in \mathcal{P}(\mathbb{N})$.

(a) Prove that $H(P) \leq \log(\frac{\pi^2}{6}) + 2\mathbb{E}_P[\log(X)]$.

Solution. We use the fact that $q(n) = \frac{6}{\pi^2 n^2}$ is a valid PMF on $\mathbb{N}$. This is in part due to the fact that $\sum_{i=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} = \zeta(2)$. That result could also be obtained from Euler's infinite product for $sin(x)$.

$$\begin{aligned} H(P) &= \mathbb{E}_P\left[\log \frac{1}{p(\mathbb{N})}\right] \\ &= \sum_x P(X=x) \log \frac{1}{P(X=x)} = -\sum_x P(X=x) \log P(X=x) \end{aligned}$$

Consider $D_{\mathsf{KL}}\left(P \| Q\right) = \sum p(x) \log \frac{p(x)}{q(x)} = \sum p(x)(\log p(x) - \log q(x))$. Note we know that divergences are non-negative, so $0 \leq D_{\mathsf{KL}}\left(P \| Q\right)$, which gives us

$$\begin{aligned} -\sum p(x) \log p(x) &\leq -\sum p(x) \log q(x) \\ &= -\sum p(x) \log \frac{6}{\pi^2 x^2} \\ &= \log(\frac{\pi^2}{6}) + 2\sum p(x) \log x \\ &= \log(\frac{\pi^2}{6}) + 2\mathbb{E}_P[\log X] \end{aligned}$$

■

(b) Provide an example of a distribution of P such that $H(P) = \infty$.

Solution. Given the above, a distribution that has infinite shannon entropy must have $\mathbb{E}_P[\log X] = \infty$. Let $p(n) = \frac{c}{n \log^2 n}$ (for $n \geq 2$ and 0 otherwise) where c is some normalization constant. We know that the sum $\sum\limits_{n=2}^{\infty} \frac{1}{n \log^2 n}$ converges from [1] so it is a valid pmf.
Note that $\mathbb{E}_P[\log X] = \sum p(x) \log x = \frac{\log x}{x \log^2 x} = \frac{1}{x \log x}$ which diverges. Now we need to confirm that the shannon entropy also diverges.

$$\begin{aligned} H(P) &= \sum -p(x) \log p(x) \\ &= \sum p(x) \log(x \log^2(x)) \\ &= \sum p(x) \log x + \sum p(x) \log^2 x \\ &= \mathbb{E}_P[\log X] + \sum p(x) \log^2 x && \text{note that the first term diverges} \\ &= \infty \end{aligned}$$

■

[1] https://math.stackexchange.com/questions/574503/infinite-series-sum-n-2-infty-frac1n-log-n

6. **Convexity/Concavity of Mutual Information.**
For $(X, Y) \sim P_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ the mutual information $I(X;Y)$ is a functional of $P_{X,Y}$. With the decomposition $P_{X,Y} = P_X P_{Y|X}$, the mutual information can be equivalently represented as a functional of the pair $(P_X, P_{Y|X})$. In this question we focus on the latter representation, and henceforth use the notation $I(P_X, P_{Y|X})$ in place of $I(X;Y)$. Prove the following:

(a) For fixed $P_X, I(P_X, P_{Y|X})$ is convex in $P_{Y|X}$.

Solution. Note that $I(P_X; P_{Y|X}) = D_{\mathsf{KL}}\left(P_Y \| P_{Y|X} | P_X\right)$. Since P_X is fixed, we use the convexity of the KL Divergence to get that $D_{\mathsf{KL}}\left(P_Y \| P_{Y|X} | P_X\right)$ is convex in P_Y and $P_{Y|X}$ which gives us the desire result that $I(P_X; P_{Y|X})$ is convex in $P_{Y|X}$

■

(b) For fixed $P_{Y|X}, I(P_X, P_{Y|X})$ is concave in P_X.

Solution.

i. We have $X = P_X^{(1)}\mathbb{P}(\Theta = 1) + P_X^{(2)}\mathbb{P}(\Theta = 2) = \alpha P_X^{(1)} + (1-\alpha)P_X^{(2)}$

ii. Note that $\mathbb{P}_{Y|X}(\cdot|X) = \mathbb{P}_{Y|X}(\cdot|X, \Theta)$ since given a value $X = x$, the value of Θ only affects the distribution of X, but since X is given, then $\mathbb{P}(X = x|X = x, \Theta) = \mathbb{P}(X = x|X = x)$ as having the value of Θ would not alter the probability of X since x is given, which imply the Markov property

iii. Using the above, we have

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &\geq H(Y|\Theta) - H(Y|X) && \text{conditioning decrease entropy} \\ &= H(Y|\Theta) - H(Y|X,\Theta) && \text{follows from above} \end{aligned}$$

Note that

$$\begin{aligned} LHS &= I(\alpha P_X^{(1)} + (1-\alpha)P_X^{(2)}; P_{Y|X}) \\ &\geq I(X;Y|\Theta) \\ &= p(\Theta = 1)I(P_X^{(1)}; P_{Y|X}) + p(\Theta = 2)I(P_X^{(2)}; P_{Y|X}) \\ &= \alpha I(P_X^{(1)}; P_{Y|X}) + (1-\alpha)I(P_X^{(2)}; P_{Y|X}) \end{aligned}$$

which is the convexity result that we want.

■

7. **Mutual Information of Sums.**
Let $Z_1, Z_2, Z_3, ...$ be an i.i.d sequence of $Ber(\frac{1}{2})$ random variables. Define

$$X_i := \sum_{j=1}^{i} Z_j, \quad 1 \leq i \leq n.$$

Find $I(X_1; X_2, ..., X_n)$.

Solution. We start by first noting that these variables form a markov chain $\mathbb{P}(X_i|X_{i-1}) = \mathbb{P}(X_i|X_{i-1}, ..., X_1)$. Using this, we show that $\mathbb{P}(X_1|X_2) = \mathbb{P}(X_1|X_2, ..., X_n)$.

$$\begin{aligned}
\mathbb{P}(X_1|X_2, ..., X_n) &= \frac{\mathbb{P}(X_n, ..., X_2|X_1)\mathbb{P}(X_1)}{\mathbb{P}(X_n, ..., X_2)} \\
&= \frac{\mathbb{P}(X_n, ..., X_3|X_2, X_1)\mathbb{P}(X_2|X_1)\mathbb{P}(X_1)}{\mathbb{P}(X_n, ..., X_3|X_2)\mathbb{P}(X_2)} \\
&= \frac{\mathbb{P}(X_n, ..., X_3|X_2)\mathbb{P}(X_2|X_1)\mathbb{P}(X_1)}{\mathbb{P}(X_n, ..., X_3|X_2)\mathbb{P}(X_2)} && \text{using the markov property} \\
&= \frac{\mathbb{P}(X_2|X_1)}{\mathbb{P}(X_2)}
\end{aligned}$$

This implies that $I(X_1; X_2, ...X_n) = I(X_1; X_2)$, and we use

$$I(X_1; X_2) = H(X_1) - H(X_1|X_2)$$

We know that $H(X_1) = 1$ since it's $Ber(\frac{1}{2})$. For $H(X_1|X_2)$, we have

$$P(X_1 = 0|X_2 = 0) = P(X_1 = 1|X_2 = 2) = 1$$

and

$$P(X_1 = 1|X_2 = 0) = P(X_1 = 0|X_2 = 2) = 0,$$

so

$$\begin{aligned}
H(X_1|X_2) &= P(X_1 = 1|X_2 = 1)\log\frac{1}{P(X_1 = 1|X_2 = 1)} + P(X_1 = 0|X_2 = 1)\log\frac{1}{P(X_1 = 0|X_2 = 1)} \\
&= (\frac{1}{2})^2 + (\frac{1}{2})^2 \\
&= \frac{1}{2}
\end{aligned}$$

Finally, giving us $I(X_1; X_2, ..., X_n) = I(X_1; X_2) = 1 - \frac{1}{2} = \frac{1}{2}$

■

8. **KL Divergence and L^2 Norm.**
Let $P, Q \in \mathcal{P}([0, 1])$ with PDFs p and q respectively. Assume that

$$0 < c_1 \leq p(x), q(x) < c_2 < \infty \quad \forall x \in [0, 1].$$

Show that the KL divergence is equivalent to the L_2 distance between the two PDFs. That is, $\exists k_1, k_2 \in \mathbb{R}_{>0}$ such that

$$k_1 \int \big(p(x) - q(x)\big)^2 \mathsf{d}x \leq D_{\mathsf{KL}}\left(P\|Q\right) \leq k_2 \int \big(p(x) - q(x)\big)^2 \mathsf{d}x$$

Solution. For the lower bound, we utilize the Pinsker's inequality that we proved on the previous homework

$$\frac{1}{2}\|P-Q\|^2 \leq D_{\mathrm{KL}}(P\|Q)$$

Now we need to show $\exists k_1$ s.t. $k1\int\left(p(x)-q(x)\right)^2 \mathsf{d}x \leq \frac{1}{2}(\int |p(x)-q(x)|\mathsf{d}x)^2$. First, note that $|p(x)-q(x)| \leq c_2 - c_1$, so we get

$$\begin{aligned}(p(x)-q(x))^2 &= |p(x)-q(x)||p(x)-q(x)| \\ &\leq |p(x)-q(x)|(c_2-c_1)\end{aligned}$$

and since $\frac{(p(x)-q(x))^2}{c_2-c_1} \leq |p(x)-q(x)|$, we get

$$\frac{\int(p(x)-q(x))^2\mathsf{d}x}{c_2-c_1} \leq \int |p(x)-q(x)|\mathsf{d}x$$

Now we need to consider squaring the integral on the right hand side.

$$\begin{aligned}\left(\int |p(x)-q(x)|\mathsf{d}x\right)^2 &= \left(\int |p(x)-q(x)|\mathsf{d}x\right)\left(\int |p(x)-q(x)|\mathsf{d}x\right) \\ &\leq \left(\int |p(x)-q(x)|\mathsf{d}x\right)\int_0^1 (c_2-c_1)\mathsf{d}x \\ &= \left(\int |p(x)-q(x)|\mathsf{d}x\right)(c_2-c_1)\end{aligned}$$

So finally, we have

$$\frac{\int(p(x)-q(x))^2\mathsf{d}x}{2(c_2-c_1)^2} \leq \frac{1}{2}\|P-Q\|^2 \text{[2]} \leq D_{\mathrm{KL}}(P\|Q)$$

so $k_1 = \frac{1}{2(c_2-c_1)^2}$ gives us the lower bound that we want.
For the upper bound we utilize the Taylor expansion of the logarithm.[3] We note that

$$\begin{aligned}D_{\mathsf{KL}}\left(P\|Q\right) &= \int p(x)\log\frac{p(x)}{q(x)}dx \\ &= \int p(x)\log p(x) - p(x)\log q(x)dx\end{aligned}$$

[2]This is TV distance

[3]https://math.stackexchange.com/questions/2614201/on-the-equivalence-between-the-kullback-leiber-divergence-and-the-l2-distance?answertab=oldest#tab-top

and then express $\log q(x)$ as

$$\begin{aligned}\log q(x) &= \log(q(x) - p(x) + p(x)) \\ &= \log\left(p(x)\left(\left[\frac{q(x)}{p(x)} - 1\right] + 1\right)\right) \\ &= \log p(x) + \log\left(\left[\frac{q(x)}{p(x)} - 1\right] + 1\right).\end{aligned}$$

We thus have

$$\begin{aligned}D_{\mathsf{KL}}\left(P\|Q\right) &= \int p(x)\log p(x) - p(x)\left(\log p(x) + \log\left(\left[\frac{q(x)}{p(x)} - 1\right] + 1\right)\right)\mathsf{d}x \\ &= \int -p(x)\log\left(\left[\frac{q(x)}{p(x)} - 1\right] + 1\right)\mathsf{d}x\end{aligned}$$

We now Taylor series expand the logarithm around 1 to obtain

$$\begin{aligned}\log\left(\left[\frac{q(x)}{p(x)} - 1\right] + 1\right) &= \left[\frac{q(x)}{p(x)} - 1\right] - \frac{1}{2}\left[\frac{q(x)}{p(x)} - 1\right]^2 \int_1^{\frac{q(x)}{p(x)}} \frac{1}{t^2}\mathsf{d}t \\ &= \left[\frac{q(x)}{p(x)} - 1\right] - \frac{1}{2}\left[\frac{q(x)}{p(x)} - 1\right]^2 \left(\frac{\left[\frac{q(x)}{p(x)} - 1\right]}{\left[\frac{q(x)}{p(x)}\right]}\right) \\ &= \left[\frac{q(x)}{p(x)} - 1\right] - \frac{1}{2}\left(\frac{q(x) - p(x)}{p(x)}\right)^2\left(\frac{q(x) - p(x)}{q(x)}\right).\end{aligned}$$

Plugging the above expansion into the KL-divergence yields

$$\begin{aligned}D_{\mathsf{KL}}\left(P\|Q\right) &= \int -p(x)\left(\left[\frac{q(x)}{p(x)} - 1\right] - \frac{1}{2}\left(\frac{q(x) - p(x)}{p(x)}\right)^2\left(\frac{q(x) - p(x)}{q(x)}\right)\right)\mathsf{d}x \\ &= \int p(x) - q(x)\mathsf{d}x + \frac{1}{2}\int (q(x) - p(x))^2\left[\frac{1}{p(x)} - \frac{1}{q(x)}\right]\mathsf{d}x \\ &= 0 + \frac{1}{2}\int (q(x) - p(x))^2\left[\frac{1}{p(x)} - \frac{1}{q(x)}\right]\mathsf{d}x \\ &\le \frac{1}{2}\int (q(x) - p(x))^2\left[\frac{1}{c_1} - \frac{1}{c_2}\right]\mathsf{d}x \\ &= \frac{c_2 - c_1}{2c_1c_2}\int (q(x) - p(x))^2\mathsf{d}x\end{aligned}$$

giving us the desired $k_2 = \frac{c_2 - c_1}{2c_1c_2}$ as our upper bound constant. ■