Exercise 1: The Convexity of KL Divergence

Let p and q be the PDFs of a pair of absolutely continuous distributions.

(a) Prove that the KL divergence is convex in the pair (p, q), i.e.,

$$D_{KL}(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \leq \lambda D_{KL}(p_1||q_1) + (1-\lambda)D_{KL}(p_2||q_2), \tag{1}$$

where (p_1, q_1) and (p_2, q_2) are two pairs of distributions and $0 \leq \lambda \leq 1$.
(Hint: you can use the log sum inequality, namely that $(a_1 + a_2)\log\left(\frac{a_1+a_2}{b_1+b_2}\right) \leq a_1 \log\frac{a_1}{b_1} + a_2 \log\frac{a_2}{b_2}$ holds for $a_1, a_2, b_1, b_2 \geq 0$).

Exercise 2: Mutual Information of Three Variables

Let X, Y, and Z be three discrete random variables. The mutual information of X, Y, and Z is defined as:

$$I(X;Y;Z) = \sum_x \sum_y \sum_z p(x,y,z) \log\left(\frac{p(x,y)p(x,z)p(y,z)}{p(x)p(y)p(z)p(x,y,z)}\right). \tag{2}$$

(a) Prove the lemma: $I(X;Y;Z) = I(X;Y) - H(X;Y|Z)$.

(b) Prove the following relation with the above lemma:

$$I(X;Y) = I(X;Y|Z) + I(Y;Z) - I(Y;Z|X). \tag{3}$$

Exercise 3: Smoothed Cross-Entropy Loss

Label smoothing (a.k.a. smoothed cross-entropy loss) [1] is a widely used trick in deep learning classification tasks. It can help to alleviate the "over-confidence" issue of the model and increase robustness. In the conventional cross-entropy loss, we aim to minimize the KL-divergence between d and $\pi(\mathbf{x}|\theta)$, where the ground truth distribution d is a delta-distribution (i.e., only $d_k = 1$ for the ground truth class), and $\pi(\mathbf{x}|\theta)$ is the predicted distribution by the model π parametrized by θ. The key step in label smoothing is to smooth the ground truth distribution. Specifically, given a hyper-parameter β (e.g., $\beta = 0.1$), we uniformly distribute the "energy" with the amount of β to all the g classes and reduce the "energy" of the ground truth class. Consequently, the smoothed ground truth distribution $\tilde{d}$ is

$$\tilde{d}_k = \begin{cases} \frac{\beta}{g} & \text{for } d_k = 0; \\ 1 - \beta + \frac{\beta}{g} & \text{for } d_k = 1. \end{cases} \tag{4}$$

The smoothed cross-entropy is then $D_{KL}(\tilde{d}||\pi(\mathbf{x}|\theta))$.

(a) What is the empirical risk when using the smoothed cross entropy? (Hint: some terms can be merged into a constant and ignored during implementation).

(b) How to implement the smoothed cross-entropy? We provide the signature of the function here as a reference:

```r
#' @param label ground truth vector of the form (n_samples,).
#'  Labels should be "1","2","3" and so on.
#' @param pred Predicted probabilities of the form (n_samples,n_labels)
#' @param smoothing Hyperparameter for label-smoothing

smoothed_ce_loss <- function(
label,
pred,
smoothing){
  return (loss)
}
```

References

[1] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826. 2016.