Solution 1: The Convexity of KL Divergence

(a) We expand the left side of the inequality and obtain:

$$\begin{aligned}
&D_{KL}(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \\
&= \int_{\mathcal{X}} \left((\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \right) \mathrm{d}x \\
&\leq \int_{\mathcal{X}} \left(\lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)} \right) \mathrm{d}x \\
&= \lambda \int_{\mathcal{X}} \left(p_1(x) \log \frac{p_1(x)}{q_1(x)} \right) \mathrm{d}x + (1-\lambda) \int_{\mathcal{X}} \left(p_2(x) \log \frac{p_2(x)}{q_2(x)} \right) \mathrm{d}x \\
&= \lambda D_{KL}(p_1||q_1) + (1-\lambda) D_{KL}(p_2||q_2).
\end{aligned} \tag{1}$$

Solution 2: The Mutual Information of Three Variables

(a) According to the definition of mutual information, we have

$$\begin{aligned}
&I(X;Y) - H(X;Y|Z) \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} - \sum_z \sum_x \sum_y p(z)p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \\
&= \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,y)}{p(x)p(y)} - \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,y|z)p(z)^2}{p(x|z)p(y|z)p(z)^2} \\
&= \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,y)}{p(x)p(y)} - \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,y,z)p(z)}{p(x,z)p(y,z)} \\
&= \sum_x \sum_y \sum_z p(x,y,z) \log \left(\frac{p(x,y)p(x,z)p(y,z)}{p(x)p(y)p(z)p(x,y,z)} \right) \\
&= I(X;Y;Z).
\end{aligned} \tag{2}$$

(b) Using the lemma we just proved, we obtain:

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

A recent paper [1] provides a good example of how this relation is used in the research of explainability.

Solution 3: Smoothed Cross-Entropy Loss

(a) The empirical risk is

$$\begin{aligned}
R_{\text{emp}} &= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{g} \tilde{d}_k^{(i)} \log \left(\frac{\tilde{d}_k^{(i)}}{\pi_k(\mathbf{x}^{(i)}|\theta)} \right) \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{g} \tilde{d}_k^{(i)} \log \tilde{d}_k^{(i)} - \tilde{d}_k^{(i)} \log \pi_k(\mathbf{x}^{(i)}|\theta) \right) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{g} \tilde{d}_k^{(i)} \log \pi_k(\mathbf{x}^{(i)}|\theta) + Const.
\end{aligned} \tag{4}$$

(b) The smoothed cross-entropy is implemented as follows:

```
#' @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){

  num_samples <- NROW(pred)
  num_classes<- NCOL(pred)

  # Let's make some assertions:
  #  label should be a 1-D array.one-hot encoded label is not necessary
  stopifnot(NCOL(label)==1)
  # smoothing hyperparameter in allowed range
  stopifnot((smoothing>=0 & smoothing <= 1))
  # Same amount of rows in labels and predictions
  stopifnot((NROW(label)== num_samples))
  # Predicted probabilities must have as many columns as labels
  stopifnot(length(unique(label)) == num_classes)

  #Calculate the base level
  smoothing_per_class <- smoothing / num_classes

  # build the label matrix. Shape = [ num_samples, num_classes]
  # Start with the base level
  smoothed_labels_matrix = matrix(smoothing_per_class,
                                  nrow=num_samples,ncol=num_classes)
  # Add the smoothed correct labels
  true_labels_loc=cbind(1:num_samples, label)
  smoothed_labels_matrix[true_labels_loc]= 1 - smoothing + smoothing_per_class
  cat("Labels matrix:\n")
  print(smoothed_labels_matrix)

  # Calculate the loss
  cat("Loss for each sample:\n ",
      rowSums(- smoothed_labels_matrix * log(pred)))

  loss <- mean(rowSums(- smoothed_labels_matrix * log(pred)))
  cat("\n Loss:\n",loss)

  return (loss)
}
```

```
  # Let's build a "confident model", the model has very high predicted
  #probabilities for one of the labels
  label= c(1,2,2,3,1)
  pred= rbind(
          c(0.85,0.10,0.05),
          c(0.05,0.9,0.05),
          c(0.02,0.95,0.03),
          c(0.13,0.02,0.85),
          c(0.86,0.04,0.1))
```

```r
  # cross entropy means smoothing=0
  smoothing=0
  loss<-smoothed_ce_loss(label,pred,smoothing)

## Labels matrix:
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    1    0
## [4,]    0    0    1
## [5,]    1    0    0
## Loss for each sample:
##   0.1625189 0.1053605 0.05129329 0.1625189 0.1508229
##  Loss:
##  0.1265029

  # Smoothed cross entropy
  smoothing=0.2
  loss_smooth<-smoothed_ce_loss(label,pred,smoothing)

## Labels matrix:
##            [,1]       [,2]       [,3]
## [1,] 0.86666667 0.06666667 0.06666667
## [2,] 0.06666667 0.86666667 0.06666667
## [3,] 0.06666667 0.86666667 0.06666667
## [4,] 0.06666667 0.06666667 0.86666667
## [5,] 0.86666667 0.06666667 0.06666667
## Loss for each sample:
##   0.4940709 0.4907434 0.5390262 0.537666 0.4988106
##  Loss:
##  0.5120634
```

References

[1] Rong, Yao, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. "A consistent and efficient evaluation strategy for attribution methods." In International Conference on Machine Learning, pp. 18770-18795. PMLR, 2022.