Sketch of the Proof of the Karush–Kuhn–Tucker Conditions

http://www.isi.ee.ethz.ch/teaching/courses/it1/Handout of November 9, 2016

We state the following theorem without proof:

**Theorem 1.** Let \( f(\alpha) \) be a concave function of \( \alpha = (\alpha_1, \ldots, \alpha_n) \) over the probability simplex \( \mathcal{R} \triangleq \{ \alpha : \alpha_i \geq 0 \ \forall i, \ \sum_{i=1}^{n} \alpha_i = 1 \} \). Assume that the partial derivatives, \( \frac{\partial f(\alpha)}{\partial \alpha_k} \) are defined and continuous over the simplex \( \mathcal{R} \) with the possible exception that \( \lim_{\alpha_k \downarrow 0} \frac{\partial f(\alpha)}{\partial \alpha_k} \) may be \( +\infty \).

Then, for \( \lambda' \in \mathbb{R} \),

\[
\left. \frac{\partial f(\alpha)}{\partial \alpha_k} \right|_{\alpha = \alpha^*} = \lambda' \quad \forall k \text{ such that } \alpha_k^* > 0, \tag{1}
\]

\[
\left. \frac{\partial f(\alpha)}{\partial \alpha_k} \right|_{\alpha = \alpha^*} \leq \lambda' \quad \forall k \text{ such that } \alpha_k^* = 0 \tag{2}
\]

are necessary and sufficient conditions on a probability vector \( \alpha^* \) to maximize \( f(\alpha) \) over \( \mathcal{R} \).


We now apply Theorem 1 to a DMC with transition probabilities \( W(y|x) \). For any probability mass function \( Q \), let

\[
(QW)(y) = \sum_{x \in \mathcal{X}} Q(x)W(y|x), \quad y \in \mathcal{Y}.
\]

**Theorem 2.**

a) Let \( Q^* \) be a probability mass function and let \( \lambda \) be a number. If \( Q^* \) and \( \lambda \) satisfy

\[
D(W(\cdot|x)||Q^*W(\cdot)) = \lambda \quad \forall x \in \mathcal{X} : Q^*(x) > 0 \quad \text{and} \tag{3}
\]

\[
D(W(\cdot|x)||Q^*W(\cdot)) \leq \lambda \quad \forall x \in \mathcal{X} : Q^*(x) = 0, \tag{4}
\]

then \( Q^* \) maximizes \( I(Q,W) \) and the capacity of the DMC is equal to \( \lambda \), i.e.,

\[
C = I(Q^*,W) = \lambda.
\]

b) Let \( Q^* \) be a probability mass function. If \( Q^* \) maximizes \( I(Q,W) \), i.e., if \( I(Q^*,W) = C \), then

\[
D(W(\cdot|x)||Q^*W(\cdot)) = C \quad \forall x \in \mathcal{X} : Q^*(x) > 0 \quad \text{and} \tag{5}
\]

\[
D(W(\cdot|x)||Q^*W(\cdot)) \leq C \quad \forall x \in \mathcal{X} : Q^*(x) = 0. \tag{6}
\]
Proof. We want to apply Theorem 1. The mutual information between the input and the output of a DMC can be expressed as

\[ I(Q, W) = \sum_x \sum_y Q(x)W(y|x) \log \frac{W(y|x)}{\sum_{x'} Q(x')W(y|x')} \]

and is a concave function of \( Q \). Here \( Q \) corresponds to \( \alpha \) in Theorem 1 (and consequently \( Q_k = Q(x_k) \) corresponds to \( \alpha_k \) in Theorem 1). The partial derivatives satisfy the requirements of Theorem 1. In order to find the maximum of \( I(Q, W) \) over all different choices of \( Q \), we have to calculate the derivatives \( \frac{\partial I(Q, W)}{\partial Q_k} \). Without loss of generality we assume that the logarithms are natural logarithms. There are two positions in \( I(Q, W) \) where \( Q_k \) appears: right after the double sum and inside to logarithm right after the summation over \( x' \). We thus have by the product rule and by the chain rule

\[
\frac{\partial I(Q, W)}{\partial Q_k}
= \sum_y W(y|x_k) \log \sum_{x'} Q(x')W(y|x')
+ \sum_x \sum_y Q(x)W(y|x) \cdot \sum_{x'} Q(x')W(y|x') \cdot \frac{W(y|x)}{W(y|x')} \cdot (-W(y|x)) \cdot \frac{W(y|x_k)}{(\sum_{x'} Q(x')W(y|x'))^2}
= \sum_y W(y|x_k) \log \sum_{x'} Q(x')W(y|x) - \sum_y \frac{W(y|x_k)}{W(y|x)} \sum_x Q(x)W(y|x)
= \sum_y W(y|x_k) \log \sum_{x'} Q(x')W(y|x') - \sum_y W(y|x_k)
= \sum_y W(y|x_k) \log \sum_{x'} Q(x')W(y|x') - 1
= D(W(\cdot|x_k)||QW(\cdot)) - 1.
\]

We are now ready to prove Part a). Since (3) and (4) are satisfied, we can invoke Theorem 1 (with \( \lambda' = \lambda - 1 \)) to conclude that \( Q^* \) maximizes \( I(Q, W) \). Then,

\[
C = I(Q^*, W) = \sum_x \sum_y Q^*(x)W(y|x) \log \frac{W(y|x)}{\sum_{x'} Q^*(x')W(y|x')}
= \sum_x Q^*(x)D(W(\cdot|x)||Q^*W(\cdot))
= \lambda,
\]

where the last equality follows from (3).

We finish by proving Part b). Because \( Q^* \) maximizes \( I(Q, W) \), we know by Theorem 1 that there exists a \( \lambda' \) such that

\[
D(W(\cdot|x)||Q^*W(\cdot)) = \lambda' + 1 \quad \forall x \in X: Q^*(x) > 0 \quad \text{and}
D(W(\cdot|x)||Q^*W(\cdot)) \leq \lambda' + 1 \quad \forall x \in X: Q^*(x) = 0.
\]

From the same computation as in (5) we obtain \( I(Q^*, W) = \lambda' + 1 \). Because we know that \( I(Q^*, W) = C \), it follows that \( \lambda' + 1 = C \) and the proof is complete.