Information Theory I

Fall Semester 2016

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory

Institut für Signal- und Informationsverarbeitung

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 $f(\boldsymbol{\alpha})$ be a concave function of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ over the probability simplex $\mathcal{R} \triangleq \left\{\boldsymbol{\alpha}\colon \alpha_i \geq 0\ \forall i,\ \sum_{i=1}^n \alpha_i = 1\right\}$. Assume that the partial derivatives, $\frac{\partial f(\boldsymbol{\alpha})}{\partial \alpha_k}$ are defined and continuous over the simplex $\mathcal{R}$ with the possible exception that $\lim_{\alpha_k \downarrow 0} \frac{\partial f(\boldsymbol{\alpha})}{\partial \alpha_k}$ may be $+\infty$. Then, for $\lambda' \in \mathbb{R}$,*

$$\left.\frac{\partial f(\boldsymbol{\alpha})}{\partial \alpha_k}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^\star} = \lambda' \qquad \forall\, k \text{ such that } \alpha_k^\star > 0, \tag{1}$$

$$\left.\frac{\partial f(\boldsymbol{\alpha})}{\partial \alpha_k}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^\star} \leq \lambda' \qquad \forall\, k \text{ such that } \alpha_k^\star = 0 \tag{2}$$

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

Proof. See Robert G. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, 1968, Theorem 4.4.1. □

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 $Q^\star$ be a probability mass function and let λ be a number. If $Q^\star$ and λ satisfy

$$D\big(W(\cdot|x) \| (Q^\star W)(\cdot)\big) = \lambda \qquad \forall x \in \mathcal{X} : Q^\star(x) > 0 \quad \textit{and} \tag{3}$$

$$D\big(W(\cdot|x) \| (Q^\star W)(\cdot)\big) \leq \lambda \qquad \forall x \in \mathcal{X} : Q^\star(x) = 0, \tag{4}$$

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

$$\mathsf{C} = I(Q^\star, W) = \lambda.$$

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

$$D\big(W(\cdot|x) \| (Q^\star W)(\cdot)\big) = \mathsf{C} \qquad \forall x \in \mathcal{X} : Q^\star(x) > 0 \quad \textit{and}$$

$$D\big(W(\cdot|x) \| (Q^\star W)(\cdot)\big) \leq \mathsf{C} \qquad \forall x \in \mathcal{X} : Q^\star(x) = 0.$$

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 $\boldsymbol{\alpha}$ in Theorem 1 (and consequently $Q_k = Q(x_k)$ corresponds to α_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

$$\begin{aligned}
&\frac{\partial I(Q,W)}{\partial Q_k} \\
&\quad= \sum_y W(y|x_k) \log \frac{W(y|x_k)}{\sum_{x'} Q(x')W(y|x')} \\
&\qquad+ \sum_x \sum_y Q(x)W(y|x) \cdot \frac{\sum_{x'} Q(x')W(y|x')}{W(y|x)} \cdot \bigl(-W(y|x)\bigr) \cdot \frac{W(y|x_k)}{\bigl(\sum_{x'} Q(x')W(y|x')\bigr)^2} \\
&\quad= \sum_y W(y|x_k) \log \frac{W(y|x_k)}{\sum_{x'} Q(x')W(y|x')} - \sum_y \frac{W(y|x_k)}{\sum_{x'} Q(x')W(y|x')} \sum_x Q(x)W(y|x) \\
&\quad= \sum_y W(y|x_k) \log \frac{W(y|x_k)}{\sum_{x'} Q(x')W(y|x')} - \sum_y W(y|x_k) \\
&\quad= \sum_y W(y|x_k) \log \frac{W(y|x_k)}{\sum_{x'} Q(x')W(y|x')} - 1 \\
&\quad= D\bigl(W(\cdot|x_k)||(QW)(\cdot)\bigr) - 1.
\end{aligned}$$

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 $Q^\star$ maximizes $I(Q,W)$. Then,

$$\begin{aligned}
\mathsf{C} = I(Q^\star, W) &= \sum_x \sum_y Q^\star(x)W(y|x) \log \frac{W(y|x)}{\sum_{x'} Q^\star(x')W(y|x')} \\
&= \sum_x Q^\star(x) D\bigl(W(\cdot|x)||(Q^\star W)(\cdot)\bigr) \\
&= \lambda, \qquad (5)
\end{aligned}$$

where the last equality follows from (3).

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

$$\begin{aligned}
D\bigl(W(\cdot|x)||(Q^\star W)(\cdot)\bigr) &= \lambda' + 1 \qquad \forall x \in \mathcal{X} : Q^\star(x) > 0 \quad \text{and} \\
D\bigl(W(\cdot|x)||(Q^\star W)(\cdot)\bigr) &\le \lambda' + 1 \qquad \forall x \in \mathcal{X} : Q^\star(x) = 0.
\end{aligned}$$

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