An Algorithm for Computing the Shannon-Capacity of Arbitrary Channels

Jürgen Krob
University of Kaiserslautern

Holger R. Scholl
University of Kaiserslautern

October 11, 1993

Abstract

Questions arising from Statistical Decision Theory, Bayes Methods and other probability theoretic fields lead to concepts of orthogonality of a family of probability measures. In this paper we therefore give a sketch of a generalized information theory which is very helpful in considering and answering those questions.
In this adapted information theory Shannon's classical transition channels modelled by finite stochastic matrices are replaced by compact families of probability measures that are uniformly integrable. These channels are characterized by concepts such as information rate and capacity and by optimal priors and the optimal mixture distribution.
For practical studies we introduce an algorithm to calculate the capacity of the whole probability family which is applicable even for general output space. We then explain how the algorithm works and compare its numerical costs with those of the classical Arimoto–Blahut–algorithm.

Keywords: Families of Probability Measures, Information Theory, Optimal Prior Distribution, Shannon-Capacity, Statistical Experiments, Structure Theory

1 Introduction

Let $\mathcal{E} = (X, \mathcal{X}, (P_\vartheta)_{\vartheta \in \Theta})$ be a dominated statistical experiment. We can identify $(P_\vartheta)_{\vartheta \in \Theta}$ with the set of densities $(f_\vartheta)_{\vartheta \in \Theta}$ with respect to the dominating measure μ which we can assume to be a probability measure. By considering the question 'How informative is a statistical experiment' several concepts have been developed to describe the structure of the family of probability measures $(P_\vartheta)_{\vartheta \in \Theta}$. Most approaches to this problem utilize pairwise comparison of the P_ϑ. One important tool of this class is the *Hellinger-metric*. Other concepts well known in statistical decision theory are based on the *Kullback-Leibler-distance* or the *Fisher-information* (For details see for instance [11], [12] and many more). But with most of these tools either it is hard to study more than local pairwise effects or they depend on the particular parametrization.

That is why new ideas have been developed to describe the global structure of $(P_\vartheta)_{\vartheta \in \Theta}$ in a way which is independent of a particular parametrization. One of these methods is based on the information theory introduced in 1949 by C.E. SHANNON (see [18]). Shannon describes a transition channel in a probabalistic way by specifying two finite sets I and O together with a stochastic matrix $\mathcal{P}$. I, the **input–alphabet**, consists of all characters which may be sent, and O, the **output–alphabet**, includes all characters that can be received on the output side of the channel. The **transition–matrix** $\mathcal{P}$ contains the probabilities with that an output character may be received given the input letter that is sent.

Figure 1: Shannon's transition channel model

In classical information theory the input and output alphabet are assumed to be finite. So we have a finite transition matrix formed by a finite family of probability measures. However in the context of structure theory of families of measures we do not want to restrict ourselves to finite families of discrete distributions and so the next section gives a short summary of transferring the techniques of classical information theory to the general setting in which we want to work. We then interpret $\mathcal{P} := (P_\vartheta)_{\vartheta \in \Theta}$ to play the role of the transition matrix $\mathcal{P}$. In this more general situation sending a character means choosing a distinct distribution from the whole family by selecting a probability measure on $\mathcal{P}$ as a source. This source induces a mixture distribution on the set $O := X$ of observable output 'letters'.
Thus if we can adapt Shannon's theory to our more general case we will be able to deduce properties of the family $(P_\vartheta)_{\vartheta \in \Theta}$ by answering classical information theoretic questions like 'How large is the amount of uncertainty of the transition process?' or 'How to select the input letters in order to maximize the information rate?'.

2 The Capacity of Statistical Experiments

Let X be an arbitrary set and $\mathcal{X}$ a σ–field over X. We denote the set of probability measures over X by Prob$X$. We can make this set into a normed space by using the norm induced by the metric of total variation. This also provides us with a topology. Given a compact subset $\mathcal{P} := (P_\vartheta)_{\vartheta\in\Theta} \subseteq \text{Prob}X$ we shall call $\mathcal{E} = (X, \mathcal{X}, (P_\vartheta)_{\vartheta\in\Theta})$ a **compact statistical experiment**. If all the P_ϑ are absolutely continuous with respect to a measure $\mu \in \text{Prob}X$ we will study the family of densities $(f_\vartheta)_{\vartheta\in\Theta}$ instead. Therefore if $\lambda \in \text{Prob}\mathcal{P}$ is a source or **prior distribution**, the mixture distribution P_λ has density $f_\lambda = \int_{\mathcal{P}} f_\vartheta d\lambda$. We point out that we always consider the space of all sources Prob$\mathcal{P}$ with respect to the weak topology.

From now on we will assume that the family $(f_\lambda \log f_\lambda)_{\lambda\in\text{Prob}\mathcal{P}}$ is uniformly integrable. This is, for instance, always the case if X is a finite set.

A central idea in information theory is to measure the amount of uncertainty caused by a source distribution λ in terms of its **entropy**:

$$\mathcal{H}(\lambda) := -\int_X f_\lambda \log f_\lambda d\mu.$$

The expected entropy induced by the transition kernel $\mathcal{P}$ given a source λ therefore is

$$\mathbf{E}_\lambda(\mathcal{H}(P_\vartheta)) = -\int_{\mathcal{P}} \Big(\int_X f_\vartheta \log f_\vartheta d\mu\Big) d\lambda.$$

We can now formulate the basic principle of information theory:

The gain of information is proportional to the loss of entropy.

Thus we define

$$\begin{aligned}\mathcal{I}(\mathcal{E},\lambda) &:= \mathcal{H}(\lambda) - \mathbf{E}_\lambda\big(\mathcal{H}(P_\vartheta)\big) \\ &= -\int_X f_\lambda \log f_\lambda d\mu + \int_{\mathcal{P}}\int_X f_\vartheta \log f_\vartheta d\mu d\lambda\end{aligned}$$

to be the **expected information gain** or the **information rate** of λ where λ is the source. We want to point out that all the quantities in the preceeding formula are finite.

Proposition 1: *If $(f_\lambda \log f_\lambda)_{\lambda\in\text{Prob}\mathcal{P}}$ is uniformly integrable then the information rate $\mathcal{I}(\lambda) := \mathcal{I}(\mathcal{E},\lambda) : \text{Prob}\mathcal{P} \to \mathbb{R}$ is a well-defined continuous function of λ.*

Proof: Both well-definedness and continuity are implied by uniformly integrability of the family $(f_\lambda \log f_\lambda)_{\lambda\in\text{Prob}\mathcal{P}}$. For a detailed proof see [9] Theorem 4.9. Also see lemma 8 in the next section. ■

As $\mathcal{P} \subseteq \text{Prob}X$ is compact this means that Prob$\mathcal{P}$ is also compact and so the information rate has a maximum in Prob$\mathcal{P}$. We define

$$\mathcal{C}(\mathcal{P}) := \max_{\lambda\in\text{Prob}\mathcal{P}} \mathcal{I}(\lambda)$$

to be the **Shannon-capacity** of the channel $\mathcal{P}$. Obviously $\mathcal{C}$ does not depend on any source distribution and is just a function of the channel $\mathcal{P}$. Those priors $\lambda \in \text{Prob}\mathcal{P}$ achieving capacity,

$\mathcal{I}(\lambda) = \mathcal{C}(\mathcal{P})$, are called **optimal**.

Proposition 2: *Let $\mathcal{E} := (X, \mathcal{X}, \mathcal{P})$ be a compact dominated statistical experiment and $(f_\lambda \log f_\lambda)_{\lambda \in \mathrm{Prob}\mathcal{P}}$ uniformly integrable.*

1. *The set of all optimal priors is a convex subset of* $\mathrm{Prob}\mathcal{P}$.
2. *All optimal priors induce the same mixture distribution on X.*

Proof: Both 1. and 2. hold because the entropy $\mathcal{H}$ and hence $\mathcal{I}$ is a concave function. ■

We shall also call the induced output distribution optimal. The probability measures P and Q are said to be **orthogonal** if there are disjoint sets A and B with $P(A) = 1$ and $Q(B) = 1$. For example measures with disjoint supports are orthogonal in this sense and it is much easier to tell them apart than measures that are not orthogonal. In this sense $\mathcal{C}$ gives an impression of how orthogonal the family $\mathcal{P}$ is because a channel with capacity $\mathcal{C}$ just behaves like a channel formed by $2^{\mathcal{C}}$ orthogonal distributions. Further knowing optimal priors helps in using the given channel very effectively. Thus we can characterize $\mathcal{P}$ by calculating the Shannon–capacity and stating an optimal prior and output distribution. A detailed treatise of these subjects from the classical viewpoint is given in [19]. For the generalized setting see [17] and [9].

Interpreting a compact family of probability measures $\mathcal{P}$ as a transition channel in the sense of Shannon brings two main advantages in structure theory: The first is that the capacity $\mathcal{C}$ of the channel characterizes the orthogonality of the whole family rather than concepts based on pairwise comparison. But we do not lose the view for the local structure as those effects reflect in the set of optimal priors. For example we can apply the information theory in Bayessian statistics by using the shannon–optimal priors as a kind of non–subjectivistic priors in the given Bayesian problem. Remarkably the known *Jeffrey's prior* which is proportional to $\sqrt{\det I_\vartheta}$ (I_ϑ is the Fisher–information) may be shown to be asymptotically optimal in Shannon's sense (see [6]. This paper gives a good overview over information theoretic aspects in Bayesian methods.) The second advantage is that the capacity is independent of any dimensionality of the involved parameter–, input– or output–spaces. Thus we may compare experiments of different dimensions in a natural way.

Next we let $P, Q \in \mathrm{Prob}X$ with densities f and g in respect with the measure $\mu \in \mathrm{Prob}X$. We define

$$\mathcal{K}(P, Q) := \int_X f(x) \log \frac{f(x)}{g(x)} \mu(dx).$$

This **Kullback–Leibler–distance** is often referred to as the **relative entropy** as it measures how different the involved densities are. We have that $0 \leq \mathcal{K}(P, Q) \leq \infty$ and $\mathcal{K}(P, Q) = 0 \Longleftrightarrow P = Q$. However $\mathcal{K}$ is not a metric since both symmetry and the triangle–inequality fail to hold. Under regularity conditions sometimes $\mathcal{K}$ looks locally like the Hellinger–metric. There is a close relationship between the information rate and the Kullback–Leibler–distance which is expressed by the following proposition.

Proposition 3: *Let $\mathcal{E} := (X, \mathcal{X}, \mathcal{P})$ be a compact dominated statistical experiment and $(f_\lambda \log f_\lambda)_{\lambda \in \mathrm{Prob}\mathcal{P}}$ uniformly integrable. Let $\mathcal{K}$ denote the Kullback–Leibler–distance and $S_{n-1} := \{s \in \mathbb{R}^n : s_i \geq 0, \sum_{i=1}^n s_i = 1\}$ be the n–dimensional standard–simplex. Then the following holds:*

1. *Let $\lambda \in \mathrm{Prob}\mathcal{P}$ be a prior. Then*

$$\mathcal{I}(\lambda) = \int_{\mathcal{P}} \mathcal{K}(P_\vartheta, P_\lambda)\, d\lambda.$$

2. *Let $\lambda^{(1)}, \cdots, \lambda^{(N)}, \lambda \in \mathrm{Prob}\mathcal{P}$ and $s = (s_1, \cdots, s_N) \in S_{N-1}$ be a probability vector. If $\lambda^* := \sum_{i=1}^N s_i \lambda^{(i)}$ then*

$$\sum_{i=1}^N s_i \mathcal{I}(\lambda^{(i)}) - \mathcal{I}(\lambda^*) = \mathcal{K}(P_{\lambda^*}, P_\lambda) - \sum_{i=1}^N s_i \mathcal{K}(P_{\lambda^{(i)}}, P_\lambda)$$

Proof: See Korollar 4.35, Theorem 4.38 of [9] and Satz 3.5 of [17].

■

Our next aim is to compute the capacity and – at least one – optimal prior. The central tool is the following theorem which is partly by Shannon (1949) and by Eisenberg and Gallager (1962). The original proof is hard to adapt to our case of possibly continuous input– and output–spaces. However Topsøe gives an intuitive proof (see [19]) which can be transferred to our general case.

Theorem 4: *Let $\mathcal{E} := (X, \mathcal{X}, \mathcal{P})$ be a compact dominated statistical experiment and $(f_\lambda \log f_\lambda)_{\lambda \in \mathrm{Prob}\mathcal{P}}$ uniformly integrable. Let $\lambda \in \mathrm{Prob}\mathcal{P}$ be a prior. $\mathcal{K}$ denotes the Kullback–Leibler–distance. Then the following conditions are necessary and sufficent for the optimality of λ.*

There is a constant $0 \leq C < \infty$ with

1. *$\mathcal{K}(P_\vartheta, P_\lambda) \leq C$ everywhere,*
2. *$\mathcal{K}(P_\vartheta, P_\lambda) = C$ λ–almost everywhere.*

If these conditions hold then C is the channel–capacity: $C = \mathcal{C}(\mathcal{P})$.

Proof: See Abschnitt 14, Satz 3 of [19], Theorem 3.13 of [17] or Theorem 4.41 of [9].

■

By means of this theorem we can test whether a given source λ and the corresponding mixture distribution P_λ are optimal. If this is the case then we can obtain the capacity by calculating the information rate of λ. In addition the theorem also leads to another characterization of the channel–capacity in terms of the well-known Kullback–Leibler–distance:

Corollary 5: *Under the assumptions of theorem 4 the following holds:*

$$\mathcal{C} = \min_{\lambda \in \mathrm{Prob}\mathcal{P}} \max_{P_\vartheta \in \mathcal{P}} \mathcal{K}(P_\vartheta, P_\lambda)$$

■

The set of measures with finite support is a dense subset of Prob$\mathcal{P}$ with respect to the weak topology and this is why we can approximate the channel–capacity by the information rate of such measures. Let $\Lambda_e := \{\lambda_e \in \text{Prob}\mathcal{P} \ : \ |\text{supp } \lambda_e| < \infty\}$ be the set of all sources with finite support.

Theorem 6: *If $\mathcal{E} := (X, \mathcal{X}, \mathcal{P})$ is a compact dominated statistical experiment and $(f_\lambda \log f_\lambda)_{\lambda \in \text{Prob}\mathcal{P}}$ uniformly integrable then*

$$\mathcal{C} = \sup_{\lambda_e \in \Lambda_e} \mathcal{I}(\lambda_e)$$

Proof: See Theorem 4.11 of [9].

■

Moreover if X is a finite set then theorem 4 and some theory of convex sets lead to result which is stronger than might be expected. It will make the task of finding an optimal prior easier in many important cases.

Theorem 7: *Let $\mathcal{E} := (X, \mathcal{X}, \mathcal{P})$ be a compact statistical experiment over the finite set X. Then there is an optimal prior $\lambda^* \in \text{Prob}\mathcal{P}$ with finite support such that*

$$\lambda^* \in \Lambda_e \quad \textit{and} \quad \mathcal{I}(\lambda^*) = \mathcal{C}.$$

Or equivalently:

$$\mathcal{C} = \max_{\lambda_e \in \Lambda_e} \mathcal{I}(\lambda_e) = \mathcal{I}(\lambda^*)$$

Proof: As X is finite it follows that $(f_\lambda \log f_\lambda)_{\lambda \in \text{Prob}\mathcal{P}}$ is uniformly integrable and so we can apply theorem 4. For details see Theorem 3.16 of [17] and Theorem 4.33 of[9].

■

The next section will present an algorithm for computing the channel–capacity which is based on the intuitive proof of theorem 4 given by Topsøe in [19].

3 Computing the Shannon-Capacity

There are several algorithms to calculate the capacity of finite discrete channels. It is the algorithm of Arimoto & Blahut which seems to be most frequently used in applications. It was introduced in 1972 in [1] and [4]. In its classical form this algorithm is very useful for studying finite channels on finite output–sets. In this section we present an algorithm for computing the Shannon–capacity of a larger class of channels. This will clearly help us to characterize families of probability distributions. But of course there are other applications of a generalized information theory where calculating the capacity and optimal priors is of great interest.

The algorithm was developed in 1992 based on the following intuitive idea of Topsøe:
If we are given a channel and if we are able to find an 'input–character' $P_{\vartheta_{\max}} \in \mathcal{P}$ such that the relative entropy or Kullback– Leibler–distance $\mathcal{K}(P_{\vartheta_{\max}}, P_\lambda)$ is larger than that for all other $P_\vartheta \in \mathcal{P}$ then we will increase the information rate by sending $P_{\vartheta_{\max}}$ with higher probability. Of course this means putting more mass of the source λ on $P_{\vartheta_{\max}}$ than before. We will show that the introduced algorithm converges to the Shannon–capacity of $\mathcal{P}$ under reasonable assumptions which are implied by the theorems of the last section. These are mainly the integrability conditions which are automatically fulfilled for a finite set X.

First we note the following

Lemma 8: *Let $\mathcal{E} := (X, \mathcal{X}, \mathcal{P})$ be a compact dominated statistical experiment.*
If $(f_\vartheta \log f_\lambda)_{(f_\vartheta, \lambda) \in \mathcal{P} \times \mathrm{Prob}\mathcal{P}}$ is uniformly integrable then

1. *$(f_\lambda \log f_\lambda)_{\lambda \in \mathrm{Prob}\mathcal{P}}$ is uniformly integrable.*
2. *If $f_{\vartheta^{(\nu)}} \longrightarrow f_\vartheta$ in $L^1(\mu)$ and $\lambda^{(\nu)} \longrightarrow \lambda$ then*

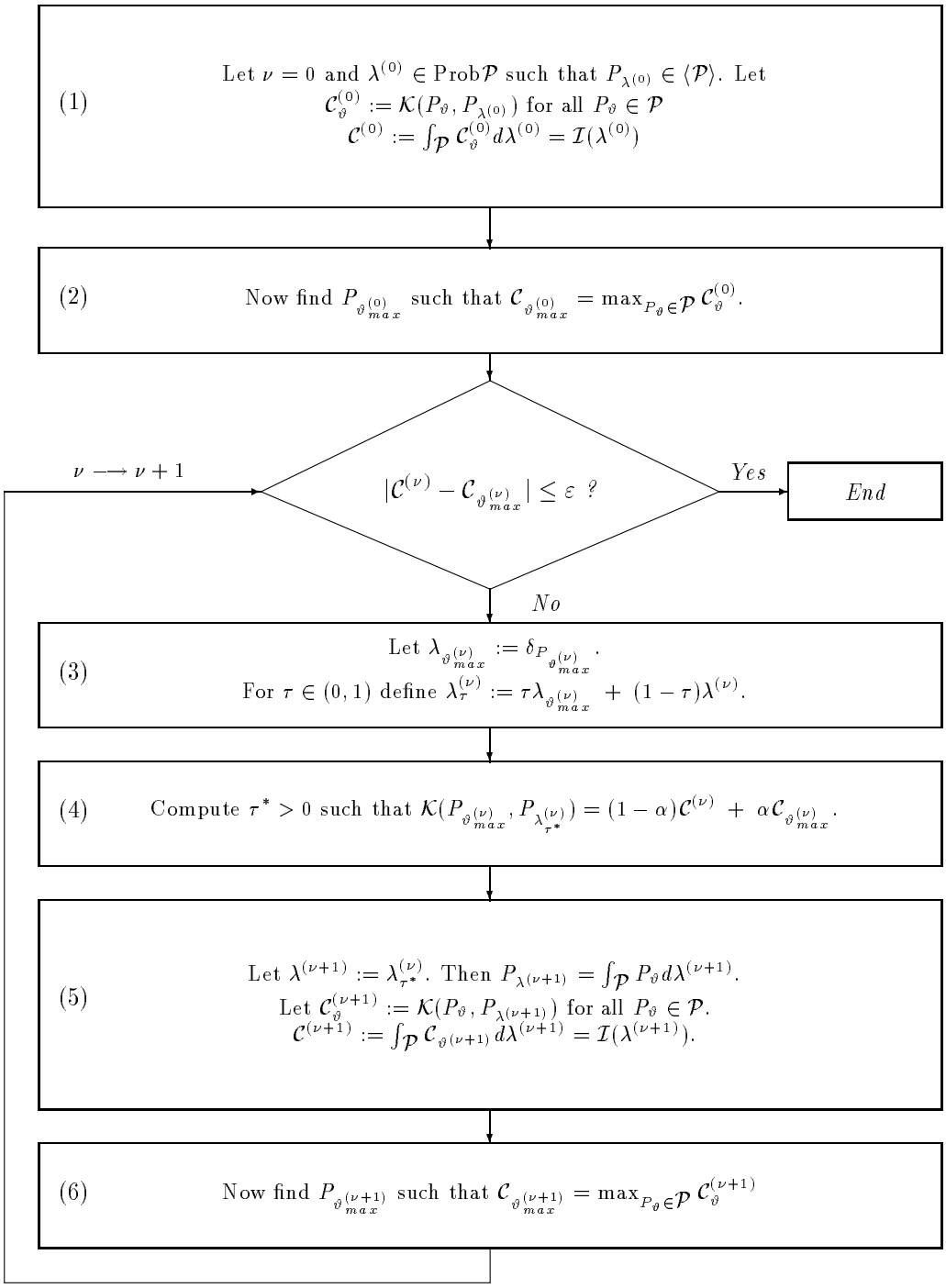
$$\mathcal{K}(P_{\vartheta^{(\nu)}}, P_{\lambda^{(\nu)}}) \overset{\nu \to \infty}{\longrightarrow} \mathcal{K}(P_\vartheta, P_\lambda).$$

Proof: See lemmata 4.46 and 4.47 in [9]. ■

The first point of the lemma means that our results thus far are applicable. The second item just means that $\mathcal{K}(\cdot, \cdot)$ is continuous in the total variation topology.

The algorithm follows on the next page.

Algorithm 9:
Let $\mathcal{E} := (X, \mathcal{X}, \mathcal{P})$ *be a compact dominated statistical experiment and* $(f_\vartheta \log f_\lambda)_{(f_\vartheta, \lambda) \in \mathcal{P} \times \mathrm{Prob}\mathcal{P}}$ *uniformly integrable. Let* $\varepsilon > 0$ *and* $0 < \alpha < 1$. *We denote the closed convex hull of* $\mathcal{P}$ *by* $\langle \mathcal{P} \rangle$.



The following theorem shows the algorithm to work correctly.

Theorem 10: *Let $\mathcal{E} := (X, \mathcal{X}, \mathcal{P})$ be a compact dominated statistical experiment and $(f_\vartheta \log f_\lambda)_{(f_\vartheta, \lambda) \in \mathcal{P} \times \mathrm{Prob}\mathcal{P}}$ uniformly integrable. Then the following holds:*

1. *The sequence of information rates computed by algorithm 9 is strictly increasing and converges to the channel–capacity.*
2. *The algorithm terminates for all $\varepsilon > 0$ in a finite number of steps.*

Proof: We give a sketch of the proof. A full proof for the case X is finite is given in [17] (see Algorithmus 4.1 and Theorem 4.4). In the cited thesis some of the assumptions are further relaxed. The general case of arbitrary X is treated in [9] (see Algorithmus 4.49 and Theorem 4.50).

Intuitively the algorithm does the following: Starting the ν^{th} step with $\lambda^{(\nu)} \in \mathrm{Prob}\mathcal{P}$ it looks for a $P_{\vartheta^{(\nu)}_{\max}} \in \mathrm{Prob}\mathcal{P}$ which has a maximum relative entropy $C_{\vartheta^{(\nu)}_{\max}}$ with respect to the mixture distribution $P_{\lambda^{(\nu)}} \in \mathrm{Prob}X$. Theorem 4 implies that if the information rate $\mathcal{I}(\lambda^{(\nu)})$ and $C_{\vartheta^{(\nu)}_{\max}}$ are equal then $\lambda^{(\nu)}$ and $P_{\lambda^{(\nu)}}$ are Shannon–optimal. If this is not the case we know that $\mathcal{I}(\lambda^{(\nu)}) < \mathcal{C}(\mathcal{P}) < C_{\vartheta^{(\nu)}_{\max}}$ but then putting a slightly larger probability on $P_{\vartheta^{(\nu)}_{\max}}$ (in steps (3) and (4)) will increase the information rate: By Proposition 3 we have

$$\begin{aligned}
\mathcal{I}(\lambda_\tau^{(\nu)}) &= \tau \underbrace{\mathcal{I}(\lambda_{\vartheta^{(\nu)}_{\max}})}_{=0} + (1-\tau)\mathcal{I}(\lambda^{(\nu)}) \;+\; \tau \mathcal{K}(P_{\vartheta^{(\nu)}_{\max}}, P_{\lambda_\tau^{(\nu)}}) \;+\; (1-\tau)\underbrace{\mathcal{K}(P_{\lambda^{(\nu)}}, P_{\lambda_\tau^{(\nu)}})}_{\geq 0} \\
&\geq (1-\tau)\mathcal{I}(\lambda^{(\nu)}) \;+\; \tau \mathcal{K}(P_{\vartheta^{(\nu)}_{\max}}, P_{\lambda_\tau^{(\nu)}})
\end{aligned}$$

As our assumptions guarantee that the Kullback–Leibler–distance is continuous we have

$$\mathcal{K}(P_{\vartheta^{(\nu)}_{max}}, P_{\tau^{(\nu)}}) \stackrel{\tau \to 0}{\longrightarrow} \mathcal{C}_{\vartheta^{(\nu)}_{max}} > \mathcal{C}^{(\nu)}.$$

But this means that there is a $\tau^* > 0$ such that

$$\mathcal{K}(P_{\vartheta^{(\nu)}_{max}}, P^{(\nu)}_{\tau^*}) = (1-\alpha)\mathcal{C}^{(\nu)} + \alpha \mathcal{C}_{\vartheta^{(\nu)}_{max}} > \mathcal{C}^{(\nu)}.$$

And with this $\tau^* > 0$ we have

$$\begin{aligned}
\mathcal{I}(\lambda^{(\nu+1)}) &= \mathcal{I}(\lambda^{(\nu)}_{\tau^*}) > (1-\tau^*)\mathcal{C}^{(\nu)} \;+\; \tau^* \mathcal{C}^{(\nu)} \\
&\stackrel{\text{Prop. 3}}{=} \mathcal{I}(\lambda^{(\nu)}),
\end{aligned}$$

which means that algorithm 9 generates an increasing sequence of information rates. This sequence is bounded by $\mathcal{C}(\mathcal{P})$ and therefore it is convergent.

Next we have to show that the limit of $\left(\mathcal{I}(\lambda^{(\nu)})\right)_{\nu \in N}$ cannot be smaller than the channel capacity. This means that it must converge to the capacity of $\mathcal{P}$. By the compactness of $\mathcal{P}$ and the Bolzano–Weierstraß–theorem we may consider the sequences $(\lambda^{(\nu)})$ and $(P_{\vartheta^{(\nu)}_{\max}})$ for $\nu \in N$ instead of convergent subsequences. Further we let λ^* and P_{λ^*} be optimal. By considering the sequences $P_{\vartheta^{(\nu)}_{\max}}$ and $P_{\tau^{(\nu)}}$ for $\nu \in N$ it can be shown by means of lemma 8 that the assumption

$$K := \lim_{\nu \to \infty} \mathcal{I}(\lambda^{(\nu)}) < \mathcal{I}(\lambda^*) = \mathcal{C}$$

leads to a contradiction. Those τ^* computed in step (4) of the algorithm per iteration ν form a sequence $(\tau^{(\nu)})_{\nu \in N}$ which can be shown to be convergent. For let $\tau^{(\nu)} \stackrel{\nu \to \infty}{\longrightarrow} \tau^*$ then by using analoguous inequalities as in the first part of the proof it follows that $\mathcal{I}(\lambda_{\tau^*})$ is a cluster point of the sequence $\left(\mathcal{I}(\lambda^{(\nu)})\right)_{\nu \in N}$ with $\tau^* > 0$. But this is impossible as this is a convergent sequence with the limit as the sole cluster point.

■

We can relax the setting of this theorem if X is a finite set. In this case the algorithm creates a sequence of mixture distributions in the interior of $\mathrm{Prob}X = S_{|X|-1}$ provided we choose a starting prior $\lambda^{(0)}$ with $P_{\lambda^{(0)}} \in \mathrm{int}\,\mathrm{Prob}X$. Lemma 8 also holds under the assumption that $\mathcal{P} \cap \mathrm{int}\mathrm{Prob}X \neq \emptyset$ and so we can prove the correctness of the algorithm even for families $\mathcal{P}$ that are not entirely included in the interior of $\mathrm{Prob}X$.