Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Summer Semester 2002 Prof. Dr. H.-A. Loeliger

ISI Internal Report

A Generalized Blahut-Arimoto Algorithm

Pascal O. Vontobel

ISI-Number: INT/200203

Abstract

Kavčić proposed in [1] an algorithm that optimizes the parameters of a Markov source at the input to a finite-state machine channel in order to maximize the mutual information rate. Numerical results for several channels indicated that his algorithm gives capacity-achieving input distributions. In this paper we prove that the stationary points of this algorithm indeed correspond one-to-one to the critical points of the information rate curve.

Kavčić's algorithm can be considered as a generalized Blahut-Arimoto algorithm, as it includes as special cases the classical Blahut-Arimoto algorithm for discrete memoryless channels and the solution to finding the capacity-achieving input distribution for finite-state channels with no noise.

Acknowledgments

First of all I would like to thank my advisor Andi Loeliger for his constant support and inspiration. A warm thanks goes also to Alek Kavčić at Harvard University, whose paper [1] was the main source of inspiration for this report. I also would like to mention my colleague Dieter Arnold with whom I had many interesting discussions on the topic of this report.

The work leading to this report was mainly done while being at the Laboratory for Signal and Information Processing at ETH Zurich. The writing of the report started in Zurich and was completed at the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign where I was supported by NSF Grant CCR 99-84515.

I'm also indebted to Steve Weller from Newcastle University in Australia for indicating some typos (Dec. 16, 2002).

Some typos were indicated to me by Alek Kavčić (Feb. 27, 2003).

Urbana, November 1, 2002 Pascal O. Vontobel

Contents

Chapter 1

Introduction

Recently, Arnold and Loeliger [2, 3], Sharma and Singh [4], and Pfister et al. [5] proposed independently a method for computing information rates of finite-state machine channels whose input is a Markov source. (These methods boil down to taking advantage of the ergodicity of the setup.) For some further historical backgroud we refer the reader to the corresponding sections in the papers just mentioned.

Subsequently, Kavčić proposed in [1] an algorithm that apparently optimizes the parameters of Markov sources at the input to a finite-state machine channel. Numerical results for several channels strongly suggested that his algorithm gives capacity-achieving Markov source parameters. The main result of this report is Th. 38 which shows that stationary points of Kavčić's algorithm indeed correspond one-to-one to the critical points of the information rate curve.

The goal of Ch. 2 is to review the classical Blahut-Arimoto algorithm for discrete memoryless channels in a way which will make the connection to the generalized Blahut-Arimoto for finite-state machine channels transparent. Ch. 3 is then devoted to Kavčić's algorithm: we show in what sense his algorithm is a generalization of the classical Blahut-Arimoto algorithm, we show that applying the generalized Blahut-Arimoto algorithm to discrete memoryless channels indeed gives the classical Blahut-Arimoto algorithm (see also [1]), and we demonstrate that applying the generalized Blahut-Arimoto algorithm to finite-state channels with no noise gives the well-known solution (see also [1]). We would like to mention that Kavčić's algorithm was also used to optimize upper bounds on the capcity of finite-state machine channels [6]. We conclude in Ch. 4 with some open problems. All proofs can be found in the corresponding appendices.

As can easily be seen from the subsequent treatment in Ch. 3, Lemma 25 is the key component that will lead to Th. 38 (see also Remark 26), and is our main contribution to the subject at hand. In our proofs we tried to be as explicit as possible, avoiding sentences like "as follows from some straightforward calculations" or "as trivally follows"; we strongly hope that this will simplify the reading. The main goal is to give a basis for discussion of further work on this topic; we are aware of the fact that Sec. 3.1 could certainly be extended to discuss all the details.

1.1 Notation

In this report we only use natural logarithms. This simplifies the proofs as we will often have to derive logarithms. Of course, all results can be formulated with respect to any other logarithm basis; the corresponding exponentiations must then of course also be modified, most notably when exponentiating $T(x)$ and T_{ij}.

We will always indicate the range of summation. The only exceptions to this rule are sums over x, y, $\mathbf{s}$, and $\mathbf{y}$. They will be sums over $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $\mathbf{s}$ over all allowed state sequences, $\mathbf{y}$ over all possible output sequences, respectively.

Chapter 2

The Blahut-Arimoto Algorithm for Discrete Memoryless Channels

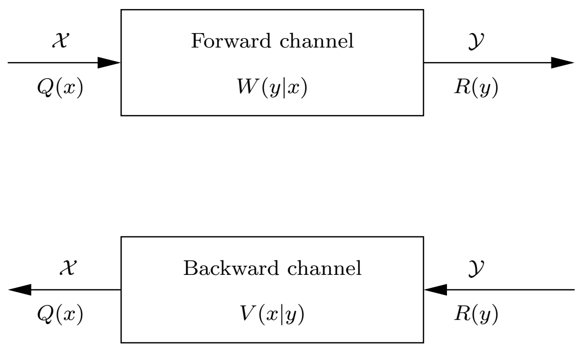

Figure 2.1: DMC with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. The "forward" channel law is given by $W(y|x)$. If the input has pmf $Q(x)$, the output has pmf $R(y) = (QW)(y)$. The "backward" channel has the channel law $V(x|y)$.

2.1 Introduction

Comment: Note that in this and the following chapters we changed the notation slightly compared with our ISIT03-Submission: The iteration number r is in $\langle\cdot\rangle$ brackets in the exponent.

The aim of this chapter is to review the Blahut-Arimoto algorithm [7, 8] (see also the tutorial [9]) for discrete memoryless channels (DMCs) in a way which will make the connection to the generalized Blahut-Arimoto for finite-state machine channels (see Chap. 3) transparent.

2.2 Discrete Memoryless Channels

Assume that we have a DMC (see Fig. 2.1) with finite input alphabet $\mathcal{X}$ and finity output alphabet $\mathcal{Y}$.[1] We assume the input to the channel to be a random variable X and $Q(x) =$

[1]The content of this chapter can easily be generalized to channels with output alphabet $\mathcal{Y} = \mathbb{R}$.

$P_X(x)$ let be the channel input pmf and let $W(y|x) = P_{Y|X}(y|x)$ be the channel law, i.e., the probability of receiving $Y = y$ when sending $X = x$. The channel output Y is also a random variable with pmf

$$P_Y(y) = R(y) \triangleq (QW)(y) \triangleq \sum_x W(y|x)Q(x). \tag{2.1}$$

The a-posteriori probability of $X = x$ upon observing $Y = y$ shall be denoted by

$$P_{X|Y}(x|y) = V(x|y) \qquad \triangleq \frac{W(y|x)Q(x)}{R(y)} \triangleq \frac{W(y|x)Q(x)}{(QW)(y)} = \frac{W(y|x)Q(x)}{\sum_{x'} Q(x')W(y|x')} \tag{2.2}$$

Because $Q(x) = \sum_y R(y)V(x|y)$, $V(x|y)$ can be considered as a "backward" channel law (see Fig. 2.1). The joint density of X and Y is therefore

$$P_{X,Y}(x,y) = Q(x)W(y|x) = R(y)V(x|y). \tag{2.3}$$

From this follows also the important relationship

$$\frac{V(x|y)}{Q(x)} = \frac{W(y|x)}{R(y)} \tag{2.4}$$

between $Q(x)$, $W(y|x)$, $R(y)$, and $V(x|y)$.

In the following, we will assume that the channel law $W(y|x)$ is *fixed*, whereas the channel input distribution $Q(x)$ will be *varied*. But note that varying $Q(x)$ will of course imply that also $R(y)$ and $V(x|y)$ *vary*! In other words, with a pmf $R(y)$ and a conditional pmf $V(x|y)$ there is implicitely a pmf $Q(x)$ behind them. Usually, we will try to make this clear by using some decorations of R and V. So, if the input pmf of X is $\tilde{Q}(x)$, then we denote the pmf of Y by $\tilde{R}(\cdot)$ and the a-posteriori probability of $X = x$ upon observing $Y = y$ is called $\tilde{V}(x|y)$. We have the relations

$$\tilde{R}(y) \triangleq (\tilde{Q}W)(y) \triangleq \sum_x \tilde{Q}(x)W(y|x), \tag{2.5}$$

$$\tilde{V}(x|y) \triangleq \frac{W(y|x)\tilde{Q}(x)}{\tilde{R}(y)} = \frac{W(y|x)\tilde{Q}(x)}{(\tilde{Q}W)(y)} = \frac{W(y|x)\tilde{Q}(x)}{\sum_{x'} \tilde{Q}(x')W(y|x')}, \tag{2.6}$$

$$\tilde{Q}(x)W(y|x) = \tilde{R}(y)\tilde{V}(x|y). \tag{2.7}$$

Note that always $\tilde{W}(y|x) = W(y|x)$, as the channel does not change (by definition), but $\tilde{V}(x|y) \neq V(x|y)$ in general.

Definition 1 (Set $\mathcal{Q}$) We let $\mathcal{Q}$ be the set of all pmfs over $\mathcal{X}$, i.e.,

$$\mathcal{Q} \triangleq \big\{Q : \mathcal{X} \to \mathbb{R} \,\big|\, Q(x) \geq 0 \text{ for all } x \in \mathcal{X}, \sum_{x \in \mathcal{X}} Q(x) = 1\big\}. \tag{2.8}$$

2.3 Mutual Information and Capacity

Definition 2 (Mutual Information) Let X and Y have the joint pmf $P_{XY}(x,y) = Q(x)W(y|x)$. The mutual information between X and Y is

$$I(Q,W) \triangleq I(X;Y) \triangleq H(Y) - H(Y|X) = H(X) - H(X|Y) \tag{2.9}$$

$$= \sum_{x,y} Q(x)W(y|x)\log\left(\frac{W(y|x)}{(QW)(y)}\right) = \sum_{x,y} Q(x)W(y|x)\log\left(\frac{V(x|y)}{Q(x)}\right). \tag{2.10}$$

Definition 3 (Channel Capacity) Let the DMC with input X and output Y have the channel law $W(y|x)$. The channel capacity is then

$$C(W) \overset{\triangle}{=} \max_{Q \in \mathcal{Q}} I(Q, W). \tag{2.11}$$

A pmf $Q \in \mathcal{Q}$ that maximizes $I(Q, W)$ is called a capacity-achieving input distributions. There are DMCs where there is not a unique capacity-achieving input distribution.

Definition 4 (Various Help Functions) We define the functions

$$f_1(Q) \overset{\triangle}{=} H(X) = -\sum_x Q(x) \log Q(x), \tag{2.12}$$

$$f_2(Q, W) \overset{\triangle}{=} H(Y) = -\sum_y (QW)(y) \log \big((QW)(y)\big), \tag{2.13}$$

$$f_3(Q, W) \overset{\triangle}{=} H(Y|X) = -\sum_x Q(x) \sum_y W(y|x) \log \left(W(y|x)\right), \tag{2.14}$$

$$f_4(Q, W) \overset{\triangle}{=} H(X|Y) = -\sum_x Q(x) \sum_y W(y|x) \log \left(\frac{Q(x)W(y|x)}{(QW)(y)}\right). \tag{2.15}$$

With these definitions we have

$$I(Q, W) = f_1(Q) - f_4(Q, W) = f_2(Q, W) - f_3(Q, W). \tag{2.16}$$

Lemma 5 *For a fixed channel law $W(y|x)$ the functions $f_1(Q)$, $f_2(Q, W)$, $f_3(Q, W)$, $f_4(Q, W)$, $I(Q, W)$ are concave in $Q(\cdot)$. Because $f_3(Q, W)$ is linear in $Q(\cdot)$, it is both concave and convex in $Q(\cdot)$.*

Proof: See Sec. A.1. □

As we will see, the classical Blahut-Arimoto algorithm strongly depends on the concavity of $I(X; Y)$ and $H(X|Y)$ as functions of $Q(\cdot)$.

2.4 The Main Idea Behind the Blahut-Arimoto Algorithm

The classical Blahut-Arimoto algorithm [7, 8, 9] solves the problem of finding numerically and in an efficent way a capacity-achieving input distribution of a DMC and therefore also the capacity of the DMC.

Fig. 2.2 schematically depicts a possible $I(Q, W)$ in function of Q. As the alphabet size $\mathcal{X}$ is usually at least two, the optimization problem is a multidimensional one. But for illustrational purposes, a one-dimensional representation of Q will do. The problem of finding a capacity-achieving input distribution is therefore to find where $I(Q, W)$ has a maximum. The problem is simplified by the fact that $I(Q, W)$ is concave in Q (as shown in Lemma 5).

There are of course different ways to find such a maximum. One of them would be to set the gradient of $I(Q, W)$ equal to zero; but this usually results in a highly non-linear equation system (see more on this in Sec. 2.6). Gradient-based methods would also lead to our goal. But a particularly elegant and efficient way to solve the problem at hand is the classical

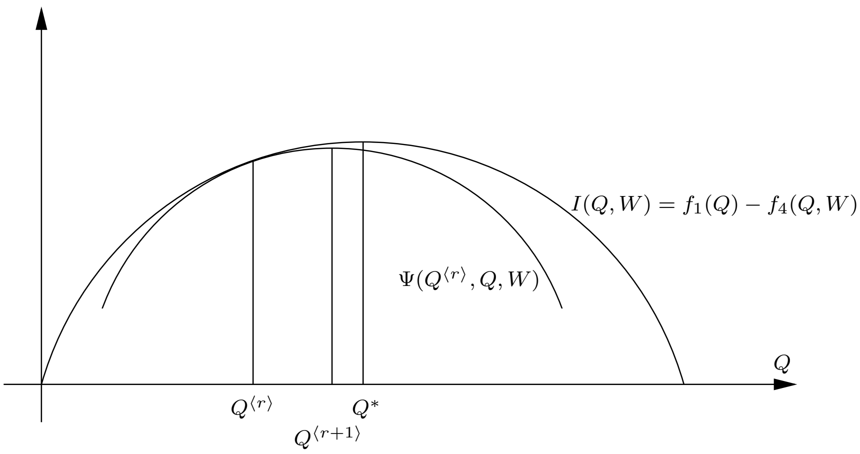

Figure 2.2: Generic mutual information $I(Q,W)$ and approximating function $\Psi(Q^{\langle r\rangle},Q,W)$. Q^* is a capacity-achieving input distribution.

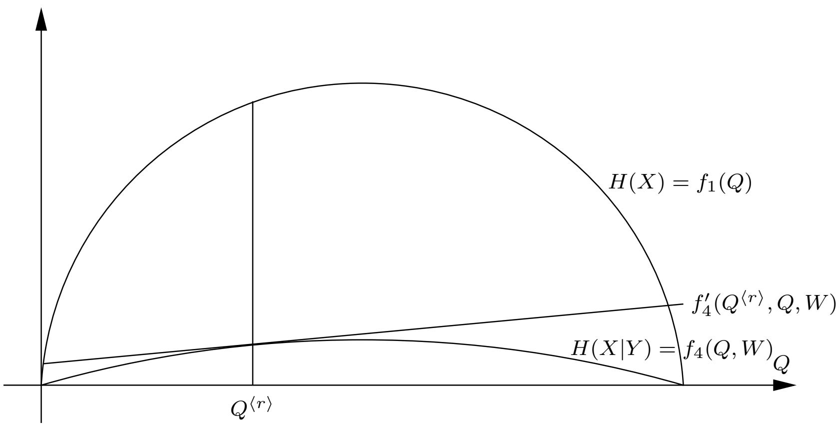

Figure 2.3: Generic entropy $H(X)$ and conditional entropy $H(X|Y)$. $f_4'(Q^{\langle r\rangle},Q,W)$ is a linear approximation of $H(X|Y)$ at $Q = Q^{\langle r\rangle}$.

Blahut-Arimoto algorithm. As it is a "nice" algorithm, there are many ways to describe it; we will choose a describe that will ease the transition to the generalized Blahut-Arimoto algorithm in Ch. 3.

The main idea of the classical Blahut-Arimoto algorithm is the following. It is an iterative algorithm, so assume that at iteration r we found some input pmf $Q^{\langle r\rangle}$ with corresponding information rate $I(Q^{\langle r\rangle}, W)$ (see Fig. 2.2). At iteration $r+1$ we would like to find a "better" $Q^{(r+1)}$, i.e., an input pmf for which $I(Q^{(r+1)}, W) \geq I(Q^{\langle r\rangle}, W)$ (see Fig. 2.2). To this end we introduce a help function $\Psi(Q^{\langle r\rangle}, Q, W)$ which locally (i.e. around $Q = Q^{\langle r\rangle}$) approximates $I(Q, W)$ (see Fig. 2.2). We require

- that the help function assumes the same value at $Q = Q^{\langle r\rangle}$ as $I(Q, W)$ does, i.e. $\Psi(Q^{\langle r\rangle}, Q^{\langle r\rangle}, W) = I(Q^{\langle r\rangle}, W)$, and
- that $\Psi(Q^{\langle r\rangle}, Q, W)$ is never above $I(Q, W)$.

If we find such a help function that can easily be maximized (let us call the pmf where the maximum is achieved $Q^{(r+1)}$), we get easily a new input pmf $Q^{(r+1)}$ with $I(Q^{(r+1)}, W) \geq I(Q^{\langle r\rangle}, W)$ (see Fig. 2.2).

The help function $\Psi(Q^{\langle r\rangle}, Q, W)$ that is used by the classical Blahut-Arimoto algorithm is the following. We express $I(Q, W)$ as

$$I(Q, W) = I(X, Y) = H(X) - H(X|Y) \stackrel{(*)}{=} f_1(Q) - f_4(Q, W), \tag{2.17}$$

where in equality $(*)$ we used the functions defined in Def. 4. Choosing

$$\Psi(Q^{\langle r\rangle}, Q, W) \stackrel{\triangle}{=} f_1(Q) - f_4'(Q^{\langle r\rangle}, Q, W), \tag{2.18}$$

where

- $f_4'(Q^{\langle r\rangle}, Q, W)$ assumes the same value at $Q = Q^{\langle r\rangle}$ as $f_4(Q, W)$ does, i.e. $f_4'(Q^{\langle r\rangle}, Q^{\langle r\rangle}, W) = f_4(Q^{\langle r\rangle}, W)$, and
- $f_4'(Q^{\langle r\rangle}, Q, W)$ is never below $f_4(Q, W)$, i.e. $f_4'(Q^{\langle r\rangle}, Q, W) \geq f_4(Q, W)$ for all Q,

leads to a function $\Psi(Q^{\langle r\rangle}, Q, W)$ fulfilling the desired requirements. By the concavity of $f_4(Q, W)$ (see Lemma 5), such a function $f_4'(Q^{\langle r\rangle}, Q, W)$ can be chosen to be the linear approximation of $f_4(Q, W)$ at $Q^{\langle r\rangle}$, i.e. the function that goes through $f_4(Q, W)$ at $Q = Q^{\langle r\rangle}$ and is tangential to $f_4(Q, W)$ (see Fig. 2.3). This is the approach taken by the classical Blahut-Arimoto algorithm.

Doing the above iterations repeatedly not only leads to input pmfs where the mutual information gets potentially larger at each iteration, but for $r \to \infty$ the input pmf $Q^{\langle r\rangle}$ converges to a capacity-achieving input distribution (see Theorem 11).

2.5 The Blahut-Arimoto Algorithm

After giving the main idea behind the Blahut-Arimoto for DMCs, we proceed to give the exact algorithm and its convergence proof.

Definition 6 ($T(x)$ Values) We assume to have a DMC with a fixed channel law $W(y|x)$. If the input pmf is $Q(x)$ we define

$$T(x) \triangleq \sum_y W(y|x)\log(V(x|y)) = \sum_y W(y|x)\log\left(\frac{Q(x)W(y|x)}{(QW)(y)}\right) \quad (\text{for all } x \in \mathcal{X}). \tag{2.19}$$

If a different input pmf is used, we will decorate the symbol T. E.g. if $\tilde{Q}(x)$ is the input pmf, we will have

$$\tilde{T}(x) \triangleq \sum_y W(y|x)\log(\tilde{V}(x|y)) = \sum_y W(y|x)\log\left(\frac{\tilde{Q}(x)W(y|x)}{(\tilde{Q}W)(y)}\right) \quad (\text{for all } x \in \mathcal{X}). \tag{2.20}$$

Definition 7 (Function Ψ) We assume to have a DMC with a fixed channel law $W(y|x)$. Let $\tilde{V}(x|y)$ for a given $\tilde{Q}(x)$ be given as in (2.6). As discussed in Sec. 2.4, the help function $\Psi(\tilde{Q}, Q, W)$ is defined as

$$\Psi(\tilde{Q}, Q, W) \triangleq -\sum_x Q(x)\log(Q(x)) + \sum_x Q(x)\sum_y W(y|x)\log(\tilde{V}(x|y)) \tag{2.21}$$

$$= \sum_x Q(x)\sum_y W(y|x)\log\left(\frac{\tilde{V}(x|y)}{Q(x)}\right) \tag{2.22}$$

$$= -\sum_x Q(x)\log(Q(x)) + \sum_x Q(x)\tilde{T}(x), \tag{2.23}$$

with

$$\tilde{T}(x) = \sum_y W(y|x)\log(\tilde{V}(x|y)) \quad (\text{for all } x \in \mathcal{X}). \tag{2.24}$$

Lemma 8 (Properties of Ψ) *For all Q, $\tilde{Q}$, and W we have*

$$\Psi\big(Q, Q, W\big) = I(Q, W) \tag{2.25}$$

and

$$\Psi\big(\tilde{Q}, Q, W\big) \le \Psi\big(Q, Q, W\big) = I(Q, W). \tag{2.26}$$

Given a W and some $\tilde{Q}$ there always exists a Q such that

$$I(\tilde{Q}, W) \le \Psi\big(\tilde{Q}, Q, W\big) \le I(Q, W). \tag{2.27}$$

Proof: See Sec. A.2 □

Remark 9 (Connection to the Outline Sec. 2.4) In the notation of Def. 7, the definitions in Sec. 2.4 are

$$\tilde{Q} = Q^{\langle r\rangle}, \tag{2.28}$$

$$f_1(Q) = -\sum_x Q(x)\log(Q(x)), \tag{2.29}$$

$$f_4(Q, W) = -\sum_x Q(x)\sum_y W(y|x)\log(V(x|y)), \tag{2.30}$$

$$f_4'(\tilde{Q}, Q, W) = -\sum_x Q(x)\sum_y W(y|x)\log(\tilde{V}(x|y)) = -\sum_x Q(x)\tilde{T}(x). \tag{2.31}$$

Note that $f_4'(\tilde{Q}, Q, W)$ is linear[2] in Q and moreover

$$f_4'(\tilde{Q}, \tilde{Q}, W) = f_4(\tilde{Q}, W), \tag{2.32}$$

$$\frac{\partial}{\partial Q(x)} f_4'(Q, W)\bigg|_{Q=\tilde{Q}} = \frac{\partial}{\partial Q(x)} f_4(Q, W)\bigg|_{Q=\tilde{Q}} = -\tilde{T}(x) \quad (\text{for all } x \in \mathcal{X}). \tag{2.33}$$

Instead of verifying the required properties as given in Sec. 2.4 of $f_4'(\tilde{Q}, Q, W)$, we have in Lemma 8 directly verified the requirements on $\Psi(\tilde{Q}, Q, W)$.

Proof: See Sec. A.3. □

Algorithm 10 (Blahut-Arimoto Algorithm for DMCs) We consider a DMC with input X and output Y and channel law $W(y|x)$. Let $Q^{\langle 0\rangle}$ be some initial (freely chosen) input distribution. For iterations $r = 0, 1, 2, \ldots$ perform alternatingly the following two steps.

- **First Step:** For each x calculate

$$T^{\langle r\rangle}(x) = \sum_y W(y|x) \cdot \log(V^{(r)}(x|y)) \tag{2.34}$$

$$= \sum_y W(y|x) \cdot \log\left(\frac{Q^{\langle r\rangle}(x)W(y|x)}{(Q^{\langle r\rangle}W)(y)}\right). \tag{2.35}$$

- **Second Step:** The new $Q^{(r+1)}(x)$ is calculated according to

$$Q^{(r+1)}(x) = \frac{e^{T^{\langle r\rangle}(x)}}{\sum_{x'} e^{T^{\langle r\rangle}(x')}}. \tag{2.36}$$

Theorem 11 (Properties of the Blahut-Arimoto Algorithm for Memoryless Channels) *For each $r = 0, 1, 2, \ldots$ the sequence of $Q^{\langle r\rangle}$ of input distributions produced by the Blahut-Arimoto algorithm fulfills*

$$I(Q^{(r+1)}, W) \geq I(Q^{\langle r\rangle}, W). \tag{2.37}$$

Furthermore, $Q^{\langle r\rangle}$ converges to a capacity-achieving input distribution for $r \to \infty$.

Proof: See Sec. A.4. See also Sec. 3.9, where the classical Blahut-Arimoto algorithm is treated as a special case of the generalized Blahut-Arimot algorithm in Alg. 37. □

Lemma 12 *Let $C = C(W)$ be the capacity for a given DMC with channel law $W(y|x)$. For any input pmf $Q(\cdot)$ we have*

$$\min_x \big[T(x) - \log(Q(x))\big] \leq I(Q, W) \leq C \leq \max_x \big[T(x) - \log(Q(x))\big], \tag{2.38}$$

[2] We could even allow an additive constant, the function would still be a linear (or, more precisely, an affine) approximation.

where $T(x)$ is as defined in Def. 6.
Proof: See Sec. A.5. □

Remark 13 (Termination Condition for Blahut-Arimoto Algorithm) From Lemma 12 we see that we can take the quantity

$$\max_x \big[T(x) - \log(Q(x))\big] - I(Q,W) \tag{2.39}$$

as a measure how close we are to capacity. We can also take the quantity

$$\max_x \big[T(x) - \log(Q(x))\big] - \min_x \big[T(x) - \log(Q(x))\big] \tag{2.40}$$

for this purpose. Note that already before the introduction of the classical Blahut-Arimoto algorithm, Gallager (see Problem 4.17 on p. 524f in [10]) proposed these capacity-achieving input distribution search termination criteria.

2.6 Intuitive Derivation of the Blahut-Arimoto Algorithm for DMCs

The aim of this section is to give another intuitive derivation of the Blahut-Arimoto algorithm as defined in Alg. 10. The idea is that one sets the gradient of $I(Q,W)$ with respect to Q equal to zero and tries to solve the resulting equations iteratively.[3]

The mutual information between X and Y is

$$I(Q,W) = \sum_x Q(x) \sum_y W(y|x) \log\left(\frac{W(y|x)}{(QW)(y)}\right) \tag{2.41}$$

$$= \sum_x Q(x) D\big(W(\cdot|x)||(QW)(\cdot)\big) \tag{2.42}$$

We would like to find the critical points of $I(Q,W)$ as a function of $Q \in \mathcal{Q}$. Neglecting in a first step the constraints $Q(x) \geq 0$, the only constraint on $Q(\cdot)$ is $\sum_x Q(x) = 1$, and we have to calculate the gradient of the Lagrangian $I(Q,W) + \lambda \sum_x Q(x)$, i.e.,

$$\frac{\partial}{\partial Q(x)} \left(I(Q,W) + \lambda \sum_x Q(x) \right) \stackrel{!}{=} 0. \tag{2.43}$$

Using the $\frac{\partial}{\partial Q(x)}(-\sum_x Q(x')\log(Q(x'))) = -\log(Q(x)) - 1$ and the results in Sec. A.3, we get

$$\sum_y W(y|x) \log\left(\frac{V(y|x)}{Q(x)}\right) - 1 + \lambda \stackrel{!}{=} 0 \tag{2.44}$$

$$\iff \sum_y W(y|x) \log\left(\frac{W(x|y)}{(QW)(y)}\right) - 1 + \lambda \stackrel{!}{=} 0 \tag{2.45}$$

$$\iff D\big(W(\cdot|x)||(QW)(\cdot)\big) - 1 + \lambda \stackrel{!}{=} 0. \tag{2.46}$$

[3] This interpretation was also mentioned in Blahut [7] (Corollary 1).

From (2.42) it follows that for a capacity-achieving input distribution $Q(\cdot)$

$$C \stackrel{!}{=} I(X;Y) = \sum_x Q(x) D\big(W(\cdot|x)||(QW)(\cdot)\big) \tag{2.47}$$

$$= \sum_x Q(x)[-\lambda + 1] = -\lambda + 1, \tag{2.48}$$

i.e., $C = -\lambda + 1$. Thus,

$$D\big(W(\cdot|x)||(QW)(\cdot)\big) = C \quad \text{(for all } x \in \mathcal{X}\text{)}. \tag{2.49}$$

(Note that we have a convex optimization problem; possibly there are different capacity achieving input distributions, but one can show that the "capacity-achieving output distribution" is unique, see also end of Sec. A.4) If we include the constraint $Q(x) \geq 0$ for all $x \in \mathcal{X}$ at the beginning we get the Kuhn-Tucker conditions

$$D\big(W(\cdot|x)||(QW)(\cdot)\big) \leq C \quad \text{(with equality if } Q(x) > 0\text{)}. \tag{2.50}$$

Formulating out the relative entropy of the Kuhn-Tucker condition we obtain (with $W(y|x)/(QW)(y) = V(x|y)/Q(x)$, $T(x) = \sum_y W(y|x) \log V(x|y)$)

$$\sum_y W(y|x) \log \left(\frac{W(y|x)}{(QW)(y)} \right) \leq C \tag{2.51}$$

$$\iff \sum_y W(y|x) \log \left(\frac{V(x|y)}{Q(x)} \right) \leq C \tag{2.52}$$

$$\iff \left(\sum_y W(y|x) \log V(x|y) \right) - \log Q(x) \leq C \tag{2.53}$$

$$\iff \log Q(x) \geq T(x) - C. \tag{2.54}$$

(Always with equality if $Q(x) > 0$.) Note that if $Q(x) = 0$, then $V(x|y) = 0$ and $T(x) = -\infty$.

From this formula we can derive the Blahut-Arimoto algorithm heuristically. If the current $Q^{\langle r \rangle}(\cdot)$–distribution does not fulfill the above equations, then one should update the density according to (for some constant $c^{\langle r \rangle}$)

$$\log Q^{(k+1)}(x) = T^{\langle r \rangle}(x) - c^{\langle r \rangle} \tag{2.55}$$

$$\iff Q^{(k+1)}(x) \propto e^{T^{\langle r \rangle}(x)} \tag{2.56}$$

$$\iff Q^{(k+1)}(x) = \frac{e^{T^{\langle r \rangle}(x)}}{\sum_{x'} e^{T^{\langle r \rangle}(x')}}, \tag{2.57}$$

where $T^{\langle r \rangle} = \sum_y W(y|x) \log V^{(r)}(x|y)$ and $V^{(r)}(x|y) = Q^{\langle r \rangle}(x) W(y|x)/((Q^{\langle r \rangle} W)(y))$. But this is exactly the Blahut-Arimoto update rule as discussed in Alg. 10. As a side result we have, that at a stationary point $e^C = \sum_x e^{T(x)}$, where C is the capacity.

Fig. 3.5 shows a trellis with one state and parallel branches from one time instant to the next bearing all the possible input letters. At each time step the question comes to with what probability one should send a letter. Apparently Eq. (2.54) gives the condition to get the best compromise between entropy and being able to decode at the receiver. $T(x)$ kind of measures the quality when one sends x: the higher the quality, the higher the probability of the use of the symbol; $T(x)$ also shows up in $e^C = \sum_x e^{T(x)}$.

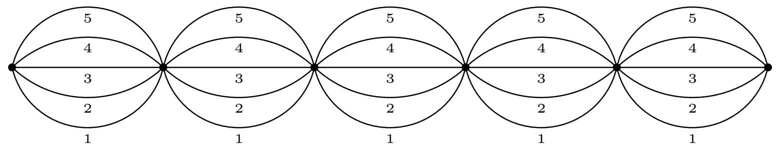

Figure 2.4: Trellis of a discrete memoryless channel with five-ary input alphabet.

Chapter 3

The Blahut-Arimoto Algorithm for Finite-State Channels

3.1 Introduction

Comment: Note that in and the following chapters we changed the notation slightly compared with our ISIT03-Submission: The iteration number r is in $\langle\cdot\rangle$ brackets in the exponent, whereas the window length parameter N is in $(\cdot)$ brackets in the exponent.

Comment: For an introduction to the subject, see the paper by Kavčić [1].

We assume that the source (channel input) is a stationary discrete-time Markov random process X_ℓ whose realizatoin x_ℓ takes on values from a finite-size source alphabet $\mathcal{X}$. It is assumed that the channel input process has memory $L \geq 0$, i.e. we have for any integer $m \geq 0$,

$$P(x_\ell|\mathbf{x}_{\ell-L-m}^{\ell-1}) = P(x_\ell|\mathbf{x}_{\ell-L}^{\ell-1}). \tag{3.1}$$

We consider an indecomposable time-independent finite-state machine channel (FSC) [10]. The channel state at time ℓ is denoted by the random variable S_ℓ, whose realization is $s_\ell \in \mathcal{S} = \{1, \ldots, M\}$. We choose the state alphabet M to be the minimum integer $M > 0$, such that s_ℓ forms a Markov process of memory 1, i.e. for any integer $m \geq 0$,

$$P(s_\ell|\mathbf{s}_{\ell-m}^{\ell-1}) = P(s_\ell|s_{\ell-1}). \tag{3.2}$$

For example, if the channel input X_ℓ is a binary Markov process of memory 3 and the channel is PR4 (i.e. $1 - D^2$) of memory 2, then $M = 2^{\max(3,2)} = 8$ guarantees that the state sequence is a Markov process of memory 1. In this report we assume that the channel is "controllable" in the sense that after a bounded number of time units we can be in a desired state. Put differently, there is (besides some bounded initial transient) a one-to-one relationship between

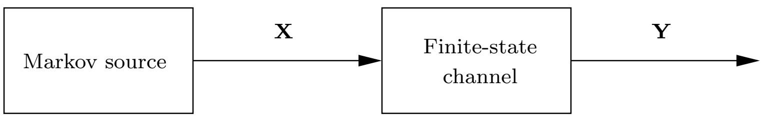

Figure 3.1: Markov source and finite-state machine channel.

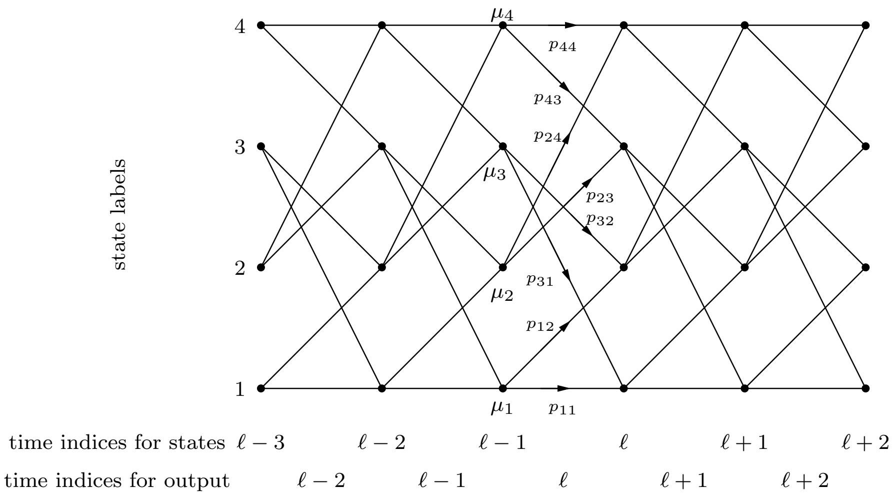

Figure 3.2: Trellis of Markov source combined with a finite-state machine channel.

input sequences $\mathbf{x}$ and state sequences $\mathbf{s}$. Channels described by a Kronecker-delta impulse response (like the PR4 channel mentioned above) are typical examples of this class.[1]

The channel output[2] $Y_\ell \in \mathcal{Y}$ is a hidden Markov sequence induced by the state sequence S_ℓ, i.e., for a discrete random variable Y_ℓ, the pmf of Y_ℓ satisfies

$$P(y_\ell|\mathbf{s}_{-\infty}^{+\infty}, \mathbf{y}_{-\infty}^{\ell-1}, \mathbf{y}_{\ell+1}^{+\infty}) = P(y_\ell|s_{\ell-1}, s_\ell). \tag{3.3}$$

For indecomposable channels, the choice of initial state does not affect the mutual information rate [10]. The following definition introduces the notation that will be used subsequently.

Definition 14 (Notation) We will use the following notation.

- The Markov process representing the FSC can be visualized by a trellis (Fig. 3.2 shows an example with $|\mathcal{S}| = 4$ states where $|\mathcal{X}| = 2$).

- Let the set $\mathcal{A}$ contain all pairs (i, j) that constitute valid state transitions. Let $\overrightarrow{\mathcal{A}_i} \triangleq \{j \,|\, (i, j) \in \mathcal{A} \text{ for some } i\}$ be the set of all valid follow-up states of state i, and let $\overleftarrow{\mathcal{A}_i} \triangleq \{k \,|\, (k, i) \in \mathcal{A} \text{ for some } i\}$ be the set of all valid preceding states of state i. (In the example in Fig. 3.2 we have $\mathcal{A} = \{(1,1), (1,2), (2,3), (2,4), (3,1), (3,2), (4,3), (4,4)\}$, $\overrightarrow{\mathcal{A}_1} = \{1, 2\}$, $\overleftarrow{\mathcal{A}_1} = \{1, 3\}$, etc.)

- If $(i, j) \in \mathcal{A}$, the transition probability of going from state i to state j is denoted by p_{ij}.

- As mentioned in the introduction, we assume that all transition probabilities are *time-independent*. Therefore it makes sense to talk about stationary state probabilities: we let μ_i be the stationary state probability of being in state $i \in \mathcal{S}$.

[1]The extension of the results of this report to more general channels where the state can only be controlled partially or not at all (e.g. the Gilbert-Elliot channel [10]) is possible, but outside the scope of this report. Moreover, here we will assume that the memory of the Markov source is at least as large as the memory of the channel.

[2]We assume here $\mathcal{Y}$ to be finite. But the results of this chapter can easily be extended to the case where $\mathcal{Y} = \mathbf{R}$.

- Let $Q_{ij} \triangleq \mu_i \cdot p_{ij}$ be the stationary probability of using the branch going from state i to branch j for $(i,j) \in \mathcal{A}$.

  If $Q_{ij} = Q_{ij}(\alpha)$ are functions of the single parameter α, we will often use the abbreviation

$$Q_{ij}^{\alpha} = Q_{ij}^{\alpha}(\alpha) \triangleq \frac{\mathrm{d}}{\mathrm{d}\alpha} Q_{ij}(\alpha). \tag{3.4}$$

- We will use the definition $N' \triangleq 2N$ and the index sets

$$\mathcal{I}_N \triangleq [-N+1, N] \qquad = \{-N+1, \ldots, N\}, \tag{3.5}$$

$$\mathcal{I}'_N \triangleq [-N+1, N-1] = \{-N+1, \ldots, N-1\}. \tag{3.6}$$

  Note that $|\mathcal{I}_N| = 2N = N'$ and that $|\mathcal{I}'_N| = 2N - 1 = N' - 1$.

- We will consider a finite window of the state and output process. For a given $N > 0$, we let the state sequence $\mathbf{s}$ be a vector with indices from $-N$ to $+N$, and the output sequence $\mathbf{y}$ be a vector with indices from $-N+1$ to $+N$. Finally, we will nearly always be interested in the limit $N \to \infty$.

- The probability of a state sequence, of an output sequence given an input sequence, of an output sequence, of a state sequence given an output sequence are, respectively,

$$Q(\mathbf{s}) \triangleq \mu_{s_{-N}} \prod_{\ell \in \mathcal{I}_N} p_{s_{\ell-1}, s_\ell} = \frac{\prod_{\ell \in \mathcal{I}_N} Q_{s_{\ell-1}, s_\ell}}{\prod_{\ell \in \mathcal{I}'_N} \sum_j Q_{s_\ell, j}}, \tag{3.7}$$

$$W(\mathbf{y}|\mathbf{s}) \triangleq \prod_{\ell \in \mathcal{I}_N} W(y_\ell | s_{\ell-1}, s_\ell), \tag{3.8}$$

$$R(\mathbf{y}) \triangleq (QW)(\mathbf{y}) \triangleq \sum_{\mathbf{s}} Q(\mathbf{s}) W(\mathbf{y}|\mathbf{s}), \tag{3.9}$$

$$V(\mathbf{s}|\mathbf{y}) \triangleq \frac{Q(\mathbf{s}) W(\mathbf{y}|\mathbf{s})}{R(\mathbf{y})} = \frac{Q(\mathbf{s}) W(\mathbf{y}|\mathbf{s})}{(QW)(\mathbf{y})} = \frac{Q(\mathbf{s}) W(\mathbf{y}|\mathbf{s})}{\sum_{\mathbf{x}'} Q(\mathbf{x}') W(\mathbf{y}|\mathbf{x}')}. \tag{3.10}$$

- We will used the following conventions. $V_\ell(i|\mathbf{y})$ will be equal to $P_{X_\ell|\mathbf{Y}}(i|\mathbf{Y})$ when one has $Q(\cdot)$ at the input. $V_{\ell-1,\ell}(i,j|\mathbf{y})$ will be equal to $P_{X_{\ell-1}X_\ell|\mathbf{Y}}(i,j|\mathbf{y})$ when one has $Q(\cdot)$ at the input. Furthermore, we use simplifications like $V_\ell(s_\ell|\mathbf{y}) = V(s_\ell|\mathbf{y})$ and $V_{\ell-1,\ell}(s_\ell|\mathbf{y}) = V(s_{\ell-1}, s_\ell|\mathbf{y})$, when the subscript is (the subscripts are) clear from the argument.

We will need a parametrization which is convenient for our purposes, i.e., over which we can easily optimize. In the expressions to come, p_{ij}, μ_i, and Q_{ij} will appear. To select the "base" parametrization, we define different manifolds and study their advantages and disadvantages.

Definition 15 (Manifold $\mathcal{P}$) We define the manifold $\mathcal{P}$ to be

$$\mathcal{P} = \left\{ (p_{ij})_{(i,j)\in\mathcal{A}} \;\middle|\; \begin{array}{l} p_{ij} \geq 0 \text{ (for all } (i,j) \in \mathcal{A}) \\ \sum_{j \in \overrightarrow{\mathcal{A}_i}} p_{ij} = 1 \text{ (for all } i \in \mathcal{S}) \end{array} \right\} \tag{3.11}$$

Definition 16 (Manifold $\mathcal{P}'$) We define the manifold $\mathcal{P}'$ to be

$$\mathcal{P}' = \left\{ \left((p_{ij})_{(i,j)\in\mathcal{A}}, (\mu_i)_{i\in\mathcal{S}}\right) \;\middle|\; \begin{array}{l} p_{ij} \geq 0 \text{ (for all } (i,j)\in\mathcal{A}) \\ \sum_{j\in\overrightarrow{\mathcal{A}}_i} p_{ij} = 1 \text{ (for all } i\in\mathcal{S}) \\ \mu_j = \sum_{i\in\mathcal{S}} \mu_i p_{ij} \text{ (for all } j\in\mathcal{S}) \end{array} \right\} \tag{3.12}$$

Definition 17 (Manifold $\mathcal{Q}$) We define the manifold $\mathcal{Q}$ to be

$$\mathcal{Q} = \left\{ (Q_{ij})_{(i,j)\in\mathcal{A}} \;\middle|\; \begin{array}{l} Q_{ij} \geq 0 \text{ (for all } (i,j)\in\mathcal{A}) \\ \sum_{(i,j)\in\mathcal{A}} Q_{ij} = 1 \\ \sum_{k\in\overleftarrow{\mathcal{A}}_i} Q_{ki} - \sum_{j\in\overrightarrow{\mathcal{A}}_i} Q_{ij} = 0 \text{ (for all } i\in\mathcal{S}) \end{array} \right\} \tag{3.13}$$

These three manifolds have these advantages and disadvantages.

- The manifold $\mathcal{P}$ is a bounded convex subset of a hyperplane, but expressing μ_i and Q_{ij} with the help of p_{ij} only is quite complicated.
- With the manifold $\mathcal{P}'$ we can express Q_{ij} easily, but the manifold itself is not a subset of a hyperplane.
- The manifold $\mathcal{Q}$ is a bounded convex subset of a hyperplane and we can derive $\{\mu_i\}$ and $\{p_{ij}\}$ easily from $\{Q_{ij}\}$:

$$\mu_i = \sum_{j\in\overrightarrow{\mathcal{A}}_i} Q_{ij} = \sum_{k\in\overleftarrow{\mathcal{A}}_i} Q_{ki}, \qquad p_{ij} = \frac{Q_{ij}}{\mu_i} = \frac{Q_{ij}}{\sum_{j'} Q_{ij'}}. \tag{3.14}$$

Because of these reasons we will choose the manifold $\mathcal{Q}$, i.e., the set Q_{ij} as our "base" parametrization. The other manifolds could also be used, but from our experience, $\mathcal{Q}$ turns out to be the most suited. So, when we are using μ_i and p_{ij}, they will always be functions of Q_{ij}, i.e.,

$$\mu_i = \sum_{j\in\overrightarrow{\mathcal{A}}_i} Q_{ij}, \qquad p_{ij} = \frac{Q_{ij}}{\mu_i} = \frac{Q_{ij}}{\sum_{j'} Q_{ij'}}. \tag{3.15}$$

When $Q_{ij} = Q_{ij}(\alpha)$ (for all $(i,j)\in\mathcal{A}$) will be a function of a single parameter α, we will implicitely also have the functions $\mu(\alpha)$ and $p_{ij}(\alpha)$. Note that to simplify notation, we will only write Q_{ij} instead of $\{Q_{ij}\}$ in function arguments, so we write $f_4(Q_{ij}, W)$ instead of $f_4(\{Q_{ij}\}, W)$, etc.

3.2 Mutual Information and Capacity

Because of the controllability assumption in Sec. 3.1, we have

$$\lim_{N\to\infty} \frac{1}{N'} I(\mathbf{X};\mathbf{Y}) = \lim_{N\to\infty} \frac{1}{N'} I(\mathbf{S};\mathbf{Y}), \tag{3.16}$$

i.e. instead of considering $\frac{1}{N'}I(\mathbf{X};\mathbf{Y})$, we will only look at $\frac{1}{N'}I(\mathbf{S};\mathbf{Y})$, also for finite N.[3]

We will also need the following functions

[3] As already mentioned in Footnote 1, this report focuses on the case where we have controllability; with some efforts, the results can be extended to the more general case.

Definition 18 (Help Functions) For a fixed channel law $W(y|x)$ the functions

$$f_1^{(N)}(Q_{ij}) \triangleq \frac{1}{N'}H(\mathbf{S}) = -\frac{1}{N'}\sum_{\mathbf{s}} Q(\mathbf{s})\log(Q(\mathbf{s})), \tag{3.17}$$

$$f_2^{(N)}(Q_{ij}, W) \triangleq \frac{1}{N'}H(\mathbf{Y}) = -\frac{1}{N'}\sum_{\mathbf{y}} (QW)(\mathbf{y})\log\big((QW)(\mathbf{y})\big), \tag{3.18}$$

$$f_3^{(N)}(Q_{ij}, W) \triangleq \frac{1}{N'}H(\mathbf{Y}|\mathbf{S}) = -\frac{1}{N'}\sum_{\mathbf{s}} Q(\mathbf{s})\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(W(\mathbf{y}|\mathbf{s})\right), \tag{3.19}$$

$$f_4^{(N)}(Q_{ij}, W) \triangleq \frac{1}{N'}H(\mathbf{S}|\mathbf{Y}) = -\frac{1}{N'}\sum_{\mathbf{s}} Q(\mathbf{s})\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(\frac{Q(\mathbf{s})W(\mathbf{y}|\mathbf{s})}{(QW)(\mathbf{y})}\right). \tag{3.20}$$

$$\tag{3.21}$$

In the limit $N \to \infty$, we define the functions $f_1(Q_{ij})$, $f_2(Q_{ij}, W)$, $f_3(Q_{ij}, W)$, $f_4(Q_{ij}, W)$ in the obvious way. If $Q_{ij} = Q_{ij}(\alpha)$ are functions of the single parameter α, we define $f_1^{(N)}(\alpha)$, $f_2^{(N)}(\alpha, W)$, $f_3^{(N)}(\alpha, W)$, $f_4^{(N)}(\alpha, W)$ and $f_1(\alpha)$, $f_2(\alpha, W)$, $f_3(\alpha, W)$, $f_4(\alpha, W)$.

Definition 19 (Mutual Information Rate) The mutual information rate we are interested in is a function of $\{Q_{ij}\}$, and W.

$$I^{(N)}(Q_{ij}, W) \triangleq \frac{1}{N'}I(\mathbf{S};\mathbf{Y}) \tag{3.22}$$

$$= f_1^{(N)}(Q_{ij}) - f_4^{(N)}(Q_{ij}, W) = \frac{1}{N'}\sum_{\mathbf{s}} Q(\mathbf{s})\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(\frac{V(\mathbf{s}|\mathbf{y})}{Q(\mathbf{s})}\right) \tag{3.23}$$

$$= f_2^{(N)}(Q_{ij}, W) - f_3^{(N)}(Q_{ij}, W) = \frac{1}{N'}\sum_{\mathbf{s}} Q(\mathbf{s})\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(\frac{W(\mathbf{y}|\mathbf{s})}{R(\mathbf{y})}\right). \tag{3.24}$$

In the limit $N \to \infty$, we define

$$I(Q_{ij}, W) \triangleq \lim_{N\to\infty} I^{(N)}(Q_{ij}, W), \tag{3.25}$$

such that

$$I^{(N)}(Q_{ij}, W) = f_1^{(N)}(Q_{ij}) - f_4^{(N)}(Q_{ij}, W) = f_2^{(N)}(Q_{ij}, W) - f_3^{(N)}(Q_{ij}, W), I(Q_{ij}, W) = f_1(Q_{ij}) - f_4(Q_{ij}, W) = \tag{3.26}$$

and hence (with the comments at the beginning of this section)

$$I(Q_{ij}, W) = \lim_{N\to\infty}\frac{1}{N'}I(\mathbf{S};\mathbf{Y}) = \lim_{N\to\infty}\frac{1}{N'}I(\mathbf{X};\mathbf{Y}). \tag{3.27}$$

If $Q_{ij} = Q_{ij}(\alpha)$ are functions of the single parameter α, we define

$$I^{(N)}(\alpha, W) \triangleq I^{(N)}(Q_{ij}(\alpha), W), \tag{3.28}$$

$$I(\alpha, W) \triangleq I(Q_{ij}(\alpha), W), \tag{3.29}$$

such that

$$I^{(N)}(\alpha, W) = f_1^{(N)}(\alpha) - f_4^{(N)}(\alpha, W) = f_2^{(N)}(\alpha, W) - f_3^{(N)}(\alpha, W), \tag{3.30}$$

$$I(\alpha, W) = f_1(\alpha) - f_4(\alpha, W) = f_2(\alpha, W) - f_3(\alpha, W). \tag{3.31}$$

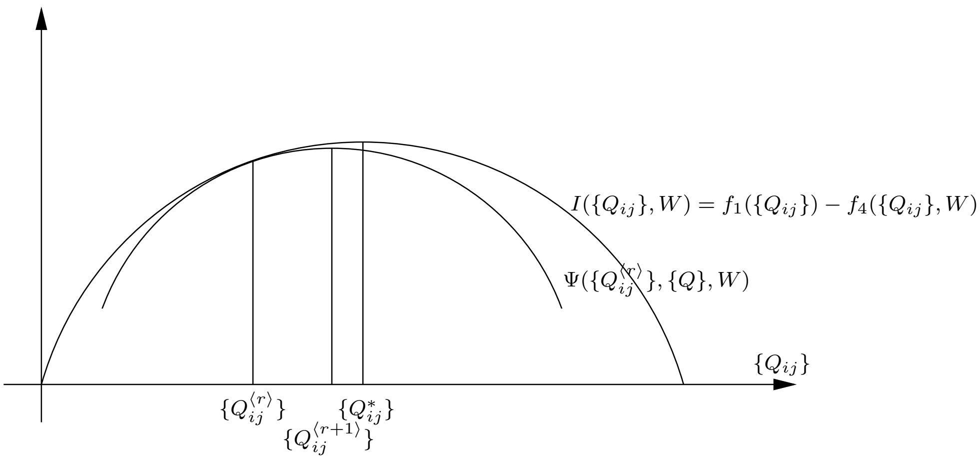

Figure 3.3: Generic mutual information rate $I(Q_{ij}, W)$ and approximating function $\Psi(Q_{ij}^{\langle r\rangle}, Q_{ij}, W)$. Q_{ij}^{*} is a capacity-achieving input branch probability distribution.

Definition 20 (Channel Capacity) With the above notation, for a given channel law W, the channel capacity for an FSC with a Markov source input is defined to be

$$C(W) \triangleq \max_{Q_{ij} \in \mathcal{Q}} I(Q_{ij}, W). \tag{3.32}$$

Gallager [10] defines the capacity slightly differently: he allows at the input to a finite-state channel any source, whereas here we assume to have a Markov source that has a certain memory length.

3.3 The Main Idea Behind the Generalized Blahut-Arimoto Algorithm for FSCs

Compared to the classical Blahut-Arimoto algorithm as shown in Ch. 2, the generalized Blahut-Arimoto algorithm for FSCs works as follows.

Again, the algorithm is of an iterative nature. Assume therefore that at iteration r we have found a set $Q_{ij}^{\langle r\rangle}$ of branch probabilities which lead to an information rate $I(Q_{ij}^{\langle r\rangle}, W)$ (see Fig. 3.3). At iteration $r+1$ we would like to find a better set $Q_{ij}^{\langle r+1\rangle}$ of branch probabilities, that lead to an information rate with $I(Q_{ij}^{\langle r+1\rangle}) \geq I(Q_{ij}^{\langle r\rangle})$ (see Fig. 3.3). Again, we introduce a help function $\Psi(Q_{ij}^{(r)}, Q_{ij}, W)$, which locally (i.e. at $Q = Q^{(r)}$) approximates (see Fig. 3.3)

$$I\big(Q_{ij}, W\big) = f_1\big(Q_{ij}\big) - f_4\big(Q_{ij}, W\big) \tag{3.33}$$

And again we get Ψ by

$$\Psi\left(Q_{ij}^{\langle r\rangle}, Q_{ij}, W\right) \triangleq f_1\big(Q_{ij}\big) - f_4'\left(Q_{ij}^{\langle r\rangle}, Q_{ij}, W\right), \tag{3.34}$$

where

$$f_4'\left(Q_{ij}^{\langle r\rangle}, Q_{ij}, W\right) \triangleq - \sum_{(i,j)\in\mathcal{A}} Q_{ij} T_{ij}^{\langle r\rangle} \tag{3.35}$$

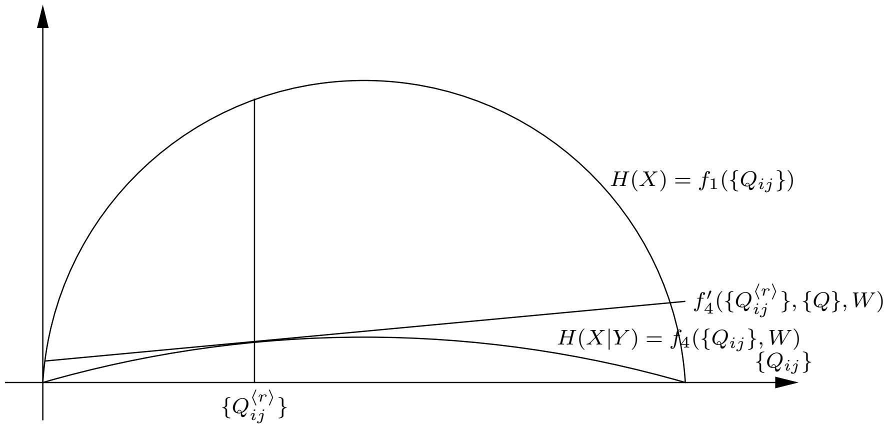

Figure 3.4: Generic entropy $H(X)$ and conditional entropy $H(X|Y)$. $f_4'(Q_{ij}^{\langle r\rangle}, Q_{ij}, W)$ is a linear approximation of $H(X|Y)$ at $Q_{ij} = Q_{ij}^{\langle r\rangle}$.

is a linear (linear in Q_{ij}) approximation of $f_4(Q_{ij}, W)$ at $\{Q_{ij}\} = \{Q_{ij}^{(r)}\}$ (see Fig. 3.4). With this approach it follows that stationary points of the algorithm correspond to zero-gradient points of the information rate curve. Moreover, zero-gradient points that are not maxima, are not stable.

Unfortunately, in this case we were neither able to show the concavity of $I(Q_{ij}, W)$, nor the concavity of $f_4(Q_{ij}, W)$ in Q_{ij}. Numerical results would suggest the truth of these hypotheses, but until now we could not found a proof or a disproof (see also Lemma 40 and Conjecture 41). The first concavity result would imply that all maxima would lie in a connected set, whereas the second concavity result would imply that the algorithm gives at each iteration a new set of branch probabilities whose associated information rate is at least as large as the old information rate.

3.4 The Generalized Blahut-Arimoto Algorithm for FSCs

The next definition introduces the crucial parameters for the generalized Blahut-Arimoto algorithm. They generalize the $T(x)$ from Chap. 2.

Definition 21 (T_{ij} values) The $T_{ij}(Q_{ij}, W)$ values[4] will be the key parameters for the generalized Blahut-Arimoto algorithm. Remember $N' = 2N$.

$$T_{ij} \triangleq \overline{\overline{T}}_{ij} - \overline{T}_i \triangleq \lim_{N\to\infty} T_{ij}^{(N)}, \qquad T_{ij}^{(N)} \triangleq \overline{\overline{T}}_{ij}^{(N)} - \overline{T}_i^{(N)}, \tag{3.36}$$

$$\begin{cases} \overline{\overline{T}}_{ij} \triangleq \lim_{N\to\infty} \overline{\overline{T}}_{ij}^{(N)}, \\ \overline{T}_i \triangleq \lim_{N\to\infty} \overline{T}_{ij}^{(N)} \end{cases} \qquad \begin{cases} \overline{\overline{T}}_{ij}^{(N)} \triangleq \frac{1}{N'} \sum_{\ell\in\mathcal{I}_N} \overline{\overline{T}}_{ij}^{(N)}(\ell) \\ \overline{T}_i^{(N)} \triangleq \frac{1}{N'} \sum_{\ell\in\mathcal{I}_N'} \overline{T}_i^{(N)}(\ell), \end{cases} \tag{3.37}$$

[4] We will often not write the arguments of the T_{ij} values; sometimes we will use decorations to indicate the arguments.

where

$$\begin{cases} \overline{\overline{T}}_{ij}^{(N)}(\ell) & \triangleq \sum\limits_{\substack{\mathbf{s} \\ s_{\ell-1}=i, s_\ell=j}} Q(\mathbf{s}_{-N}^{\ell-2}, \mathbf{s}_{\ell+1}^{N} | s_{\ell-1}, s_\ell) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_{\ell-1}, s_\ell|\mathbf{y})\right) \quad (\text{for } \ell \in \mathcal{I}_N), \\ \overline{T}_{i}^{(N)}(\ell) & \triangleq \sum\limits_{\substack{\mathbf{s} \\ s_\ell=i}} Q(\mathbf{s}_{-N}^{\ell-1}, \mathbf{s}_{\ell+1}^{N} | s_\ell) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_\ell|\mathbf{y})\right) \quad (\text{for } \ell \in \mathcal{I}'_N). \end{cases} \tag{3.38}$$

Let us mention that T_{ij} has a *dimension* in the sense that it depends on the choice of logarithm, therefore expressions like $\mathrm{e}^{T_{ij}}$ have to be modified accordingly.

Remark 39 discusses how the T_{ij} values can be calculated efficiently.

Note that we always normalize by N', despite the fact that $|\mathcal{I}'_N| = N' - 1$. (In the limit $N \to \infty$ this is irrelevant.)

Lemma 22 (Property 1 of Help Functions $f_1^{(N)}$ and f_1) *The functions $f_1^{(N)}(Q_{ij})$ and $f_1(Q_{ij})$ as given in Def. 18 can be rewritten as*

$$f_1^{(N)}(\{Q_{ij}\}) = \sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \left[-\log\left(p_{ij}\right) - \frac{1}{N'} \log\left(\mu_i\right) \right], \tag{3.39}$$

$$f_1(Q_{ij}) = \sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \left[-\log\left(p_{ij}\right) \right]. \tag{3.40}$$

Proof: See Sec. B.2. □

Lemma 23 (Property 2 of Help Functions $f_1^{(N)}$ and f_1) *Let the functions $f_1^{(N)}(\alpha)$ and $f_1(\alpha)$ be as given in Def. 18. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_1^{(N)}(\alpha) = \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \cdot \left[-\log\left(p_{ij}\right) - \frac{1}{N'} \log(\mu_i) \right], \tag{3.41}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_1(\alpha) = \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \cdot \left[-\log\left(p_{ij}\right) \right]. \tag{3.42}$$

Proof: See Sec. B.3. □

Lemma 24 (Property 1 of the Help Functions $f_4^{(N)}$ and f_4) *The functions $f_4^{(N)}(Q_{ij}, W)$ and $f_4(Q_{ij}, W)$ as given in Def. 18 can be rewritten as*

$$f_4^{(N)}(Q_{ij}, W) = -\left[\sum_{(i,j)\in\mathcal{A}} Q_{ij} \overline{\overline{T}}_{ij}^{(N)} - \sum_{i\in\mathcal{S}} \mu_i \overline{T}_i^{(N)} \right] = -\sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot T_{ij}^{(N)}, \tag{3.43}$$

$$f_4(Q_{ij}, W) = -\left[\sum_{(i,j)\in\mathcal{A}} Q_{ij} \overline{\overline{T}}_{ij} - \sum_{i\in\mathcal{S}} \mu_i \overline{T}_i \right] = -\sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot T_{ij}, \tag{3.44}$$

where $T_{ij}^{(N)} = T_{ij}^{(N)}(Q_{ij}, W)$ and $T_{ij} = T_{ij}(Q_{ij}, W)$.

Proof: See Sec. B.4. □

Lemma 25 (Property 2 of the Help Functions $f_4^{(N)}$ and f_4) *Let $Q_{ij} = Q_{ij}(\alpha)$ be functions of a single parameter α for all $(i,j) \in \mathcal{A}$, and let the functions $f_4^{(N)}(\alpha, W)$ and $f_4(\alpha, W)$ be as given in Def. 18. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_4^{(N)}(\alpha, W) = -\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \cdot T_{ij}^{(N)}, \tag{3.45}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_4(\alpha, W) = -\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \cdot T_{ij}, \tag{3.46}$$

where $T_{ij}^{(N)} = T_{ij}^{(N)}(Q_{ij}(\alpha), W)$ and $T_{ij} = T_{ij}(Q_{ij}(\alpha), W)$.

Proof: See Sec. B.5. See also Remark 26. □

Remark 26 (On Property 2 of the Help Functions $f_4^{(N)}$ and f_4) This innocent looking result in Lemma 24 is the main result of this report; the difficulty lies in the dependency of $T_{ij}^{(N)}$ and T_{ij} on α. At the point of this writing, we are not aware of an essentially shorter proof than the one shown in Sec. B.5

Definition 27 (Help Function F_4') Let $\{Q_{ij}\}, \{\tilde{Q}_{ij}\} \in \mathcal{Q}$. We define the help functions

$$f_4'^{(N)}(\tilde{Q}_{ij}, Q_{ij}, W) \triangleq -\sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \tilde{T}_{ij}^{(N)}, \tag{3.47}$$

$$f_4'(\tilde{Q}_{ij}, Q_{ij}, W) \triangleq -\sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \tilde{T}_{ij}, \tag{3.48}$$

where $\tilde{T}_{ij}^{(N)} = T_{ij}^{(N)}(\tilde{Q}_{ij}, W)$ and $\tilde{T}_{ij} = T_{ij}(\tilde{Q}_{ij}, W)$.

If $Q_{ij} = Q_{ij}(\alpha)$ are functions of a single parameter α for all $(i,j) \in \mathcal{A}$, then we define

$$f_4'^{(N)}(\tilde{\alpha}, \alpha, W) \triangleq f_4'^{(N)}(Q_{ij}(\tilde{\alpha}), Q_{ij}(\alpha), W), \tag{3.49}$$

$$f_4'(\tilde{\alpha}, \alpha, W) \triangleq f_4'(Q_{ij}(\tilde{\alpha}), Q_{ij}(\alpha), W). \tag{3.50}$$

Lemma 28 (Property 1 of the Help Functions $f_4'^{(N)}$ and f_4') *Let the functions $f_4'^{(N)}(\tilde{\alpha}, \alpha, W)$ and $f_4'(\tilde{\alpha}, \alpha, W)$ be as given in Def. 27, and let $\tilde{\alpha}$ be a fixed parameter. Then,*

$$f_4'^{(N)}(\tilde{\alpha}, \tilde{\alpha}, W) = f_4^{(N)}(\tilde{\alpha}, W), \tag{3.51}$$

$$f_4'(\tilde{\alpha}, \tilde{\alpha}, W) = f_4(\tilde{\alpha}, W), \tag{3.52}$$

Proof: See Sec. B.6. □

Lemma 29 (Property 2 of the Help Functions $f_4'^{(N)}$ and f_4') *Let the functions $f_4'^{(N)}(\tilde{\alpha}, \alpha, W)$ and $f_4'(\tilde{\alpha}, \alpha, W)$ be as given in Def. 27, and let $\tilde{\alpha}$ be a fixed parameter. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_4'^{(N)}(\tilde{\alpha}, \alpha, W)\Big|_{\alpha=\tilde{\alpha}} = f_4^{(N)}(\alpha, W)\Big|_{\alpha=\tilde{\alpha}}, \tag{3.53}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_4'(\tilde{\alpha}, \alpha, W)\Big|_{\alpha=\tilde{\alpha}} = f_4(\alpha, W)\big|_{\alpha=\tilde{\alpha}}. \tag{3.54}$$

Proof: See Sec. B.7. □

Remark 30 (On the Properties of the Help Functions $f_4'^{(N)}$ and f_4') Lemmas 28 and 29 show that $f_4'^{(N)}(\tilde{\alpha}, \alpha, W)$ is a linear approximtion of $f_4'^{(N)}(\alpha, W)$ at $\alpha = \tilde{\alpha}$, i.e., at this point they have the same value and the same gradient. The same comment applies to the functions $f_4'(\tilde{\alpha}, \alpha, W)$ and $f_4'(\alpha, W)$.

Theorem 31 (Property 1 of the Mutual Information Rate) *For any $Q_{ij} \in \mathcal{Q}$ we have*

$$I^{(N)}\left(Q_{ij}, W\right) = \sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \left[-\log\left(p_{ij}\right) - \frac{1}{N'}\log(\mu_i) + T_{ij}^{(N)}\right], \tag{3.55}$$

$$I\left(Q_{ij}, W\right) = \sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \left[-\log\left(p_{ij}\right) + T_{ij}\right], \tag{3.56}$$

where $T_{ij}^{(N)} = T_{ij}^{(N)}(Q_{ij}, W)$ and $T_{ij} = T_{ij}(Q_{ij}, W)$.

Proof: See Sec. B.8. □

Theorem 32 (Property 2 of the Mutual Information Rate) *Let $Q_{ij} \triangleq Q_{ij}(\alpha)$ for $(i,j) \in \mathcal{A}$. The derivative of $I(\alpha, W)$ with respect to α is*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} I^{(N)}(\alpha, W) = \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \cdot \left[-\log(p_{ij}) + T_{ij}^{(N)}\right], \tag{3.57}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} I(\alpha, W) = \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \cdot \left[-\log(p_{ij}) + T_{ij}\right], \tag{3.58}$$

where $T_{ij}^{(N)} = T_{ij}^{(N)}(Q_{ij}(\alpha), W)$ and $T_{ij} = T_{ij}(Q_{ij}(\alpha), W)$.

Proof: See Sec. B.9. □

Definition 33 (Generalized Function Ψ) For the generalized Blahut-Arimoto algorithm the following functions are very useful. Let

$$\Psi^{(N)}(\tilde{Q}_{ij}, Q_{ij}, W) \triangleq f_1^{(N)}(Q_{ij}) - f_4'^{(N)}(\tilde{Q}_{ij}, Q_{ij}, W), \tag{3.59}$$

$$\Psi(\tilde{Q}_{ij}, Q_{ij}, W) \triangleq f_1(Q_{ij}) - f_4'(\tilde{Q}_{ij}, Q_{ij}, W), \tag{3.60}$$

which is equivalent to

$$\Psi^{(N)}\big(\tilde{Q}_{ij}, Q_{ij}, W\big) \triangleq \sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \left[-\log(p_{ij}) - \frac{1}{N'}\log(\mu_i) + \tilde{T}_{ij}^{(N)}\right], \tag{3.61}$$

$$\Psi\big(\tilde{Q}_{ij}, Q_{ij}, W\big) \triangleq \sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \left[-\log\left(p_{ij}\right) + \tilde{T}_{ij}\right] \tag{3.62}$$

Note that the $\tilde{T}_{ij}^{(N)}$'s and $\tilde{T}_{ij}$'s are calculated according to $\tilde{Q}_{ij}$ and W, i.e. $\tilde{T}_{ij}^{(N)} = \tilde{T}_{ij}^{(N)}(Q_{ij}, W)$ and $\tilde{T}_{ij} = \tilde{T}_{ij}(Q_{ij}, W)$. When $Q_{ij} = Q_{ij}(\alpha)$ (for all $(i,j) \in \mathcal{A}$), then, as in other definitions, we define the functions $\Psi^{(N)}\big(\tilde{\alpha}, \alpha, W\big)$ and $\Psi\big(\tilde{\alpha}, \alpha, W\big)$.

Proof: See Sec. B.10. □

Theorem 34 (Property 1 of the Generalized Function Ψ) *For any $\tilde{Q}_{ij}$ we have*

$$\Psi^{(N)}\left(\tilde{Q}_{ij}, \tilde{Q}_{ij}, W\right) = I^{(N)}\left(\tilde{Q}_{ij}, W\right), \tag{3.63}$$

$$\Psi\left(\tilde{Q}_{ij}, \tilde{Q}_{ij}, W\right) = I\left(\tilde{Q}_{ij}, W\right). \tag{3.64}$$

Proof: See Sec. B.11. □

Theorem 35 (Property 2 of the Generalized Function Ψ) *Let $Q_{ij} = Q_{ij}(\alpha)$ for all $(i,j) \in \mathcal{A}$ and fix some $\tilde{\alpha}$. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\Psi^{(N)}\left(\tilde{\alpha}, \alpha, W\right)\bigg|_{\alpha=\tilde{\alpha}} = \frac{\mathrm{d}}{\mathrm{d}\alpha}I^{(N)}(\alpha, W)\bigg|_{\alpha=\tilde{\alpha}}, \tag{3.65}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\Psi\left(\tilde{\alpha}, \alpha, W\right)\bigg|_{\alpha=\tilde{\alpha}} = \frac{\mathrm{d}}{\mathrm{d}\alpha}I(\alpha, W)\bigg|_{\alpha=\tilde{\alpha}}. \tag{3.66}$$

Proof: See Sec. B.12. □

Algorithm 36 (Algorithm to Opimize Ψ) Let $\{\tilde{Q}_{ij}\}$ and W be given.[5] The $\{Q_{ij}^*\}$ defined by

$$\{Q_{ij}^*\} \triangleq \arg\max_{\{Q_{ij}\}} \Psi\left(\{\tilde{Q}_{ij}\}, \{Q_{ij}\}, W\right), \tag{3.67}$$

can be computed by the following steps.

- Let $\tilde{T}_{ij} = T_{ij}(\tilde{Q}_{ij}, W)$.
- Let $\tilde{\mathbf{A}}$ be an $|\mathcal{S}\times\mathcal{S}|$-matrix with entries $\tilde{a}_{ij} \triangleq e^{\tilde{T}_{ij}}$ if $(i,j)\in\mathcal{A}$, and $\tilde{a}_{ij} \triangleq 0$, otherwise.
- Let $\mathbf{c}$ be the left and let $\mathbf{b}^{\mathsf{T}}$ be the right eigenvector of $\tilde{\mathbf{A}}$ of the maximal (real) eigenvalue ρ of $\tilde{\mathbf{A}}$, respectively.

[5]Here we show explicitely the set braces around $\{\tilde{Q}_{ij}\}$.

- The desired solution is then

$$p_{ij}^* \triangleq \frac{b_j}{b_i} \cdot \frac{\tilde{a}_{ij}}{\rho}, \qquad \text{(for all } (i,j) \in \mathcal{A}) \tag{3.68}$$

$$\mu_i^* \triangleq K \cdot c_i \cdot b_i, \quad \text{(for all } i \in \mathcal{S}) \tag{3.69}$$

$$Q_{ij}^* \triangleq \mu_i^* \cdot p_{ij}^*, \qquad \text{(for all } (i,j) \in \mathcal{A}) \tag{3.70}$$

  where $K \triangleq 1/\left(\sum_{i \in \mathcal{S}} c_i b_i\right)$.

Moreover, the maximized value is

$$\Psi\left(\{\tilde{Q}_{ij}\}, \{Q_{ij}^*\}, W\right) = \log(\rho). \tag{3.71}$$

Proof: See Sec. B.13. □

The next algorithm was proposed by Kavčić in [1].

Algorithm 37 (Generalized Blahut-Arimoto Algorithm for FSCs) We consider a finite-state channel with state sequence $\mathbf{S}$ and output $\mathbf{Y}$ and channel law $W(\mathbf{y}|\mathbf{s})$. Let $Q_{ij}^{\langle 0 \rangle}$ be some initial (freely chosen) Markov parameter set. For iterations $r = 0, 1, 2, \ldots$ perform alternatingly the following two steps.

- **First Step:** For each $(i,j) \in \mathcal{A}$ in calculate $T_{ij}^{\langle r \rangle} = T_{ij}(\{Q_{ij}^{\langle r \rangle}\}, W)$ according to Def. 21 with Markov parameter set $\{Q_{ij}^{(r)}\}$ and channel law W. They can be approximated by the procedure given in Remark 39.

- **Second Step:** The new Markov paramter set $\{Q_{ij}^{\langle r+1 \rangle}\}$, which maximizes $\Psi(Q_{ij}^{\langle r \rangle}, Q_{ij}^{\langle r+1 \rangle}, W)$, is calculated according to Alg. 36 with inputs $\{Q_{ij}\} \triangleq \{Q_{ij}^{\langle r \rangle}\}$ and W and output $\{Q_{ij}^{\langle r+1 \rangle}\} \triangleq \{Q_{ij}^*\}$.

Theorem 38 (Properties of the Generalized Blahut-Arimoto Algorithm for FSCs) *The stationary points of Alg. 37 correspond to critical points of the information rate curve. But only local maxima of the information rate curve correspond to stable stationary points of Alg. 37.*

Further properties can be shown if the conjectures in Conj. 41 hold (see also Sec. 3.4). The concavity of $I(Q_{ij}, W)$ would imply that all maxima would lie in a connected set, whereas the concavity of $f_4(Q_{ij}, W)$ in Q_{ij} would imply that the algorithm gives at each iteration a new set of branch probabilities whose associated information rate is at least as large as the old information rate.

Proof: See Section B.14. □

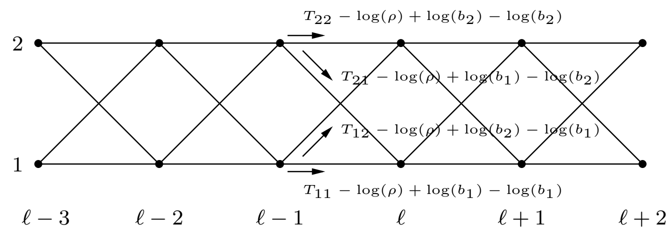

Figure 3.5: Trellis of a finite-state channel with two states. The values at the next to the arrows indicate the "gain" that one has by choosing a certain direction.

3.5 Interpretation of Stationary Points of the Generalized Blahut-Arimoto Algorithm

The update formula used in Alg 37 is

$$p_{ij}^* = \frac{b_j}{b_i} \cdot \frac{e^{T_{ij}}}{\rho}. \quad \text{(for all } (i,j) \in \mathcal{A}), \tag{3.72}$$

where we used the notation from Alg. 36.

Assume to be at a stationary point of the algorithm, i.e., $p_{ij} = p_{ij}^*$ (for all $(i,j) \in \mathcal{A}$). Logarithmically written,

$$\log(p_{ij}) = T_{ij} - \log(\rho) + \log(b_j) - \log(b_i). \tag{3.73}$$

We can give the following interpretation: T_{ij} measures the "quality" of sending the symbol when going from state i to state j. Additionally, there is potentially also an advantage to go from state i to state j, as possibly one has there a slightly better information rate: this is measured by the difference $\log(b_j) - \log(b_i)$. (Of course, asymptotically one can transmit the same normalized information, but unnormalized there is potentially a slight difference.)

Using the left eigenvector of the $\tilde{\mathbf{A}}$ matrix, we equivalently get ($\overleftarrow{p}_{ij}$ is the backward branching probability from state i to state j)

$$\log(\overleftarrow{p}_{ij}) = T_{ji} - \log(\rho) + \log(c_j) - \log(c_i). \tag{3.74}$$

We can give a similar interpretation, but now looking to the left (i.e. to the past).

For Q_{ij} we get

$$\log(Q_{ij}) = T_{ij} - \log(\rho) + \log(K) + \log(c_i) + \log(b_j), \tag{3.75}$$

where $K = 1/\left(\sum_i c_i b_i\right)$. Also here we can give some interpretation.

In (2.50) we gave the Kuhn-Tucker conditions for a capacity-achieving input pmf for a DMC, where we also took care of the case where some inputs are not used, i.e. some $Q(x)$ are zero. They can be written as

$$-\log(Q(x)) + T(x) \leq C \quad \text{(with equality if } Q(x) > 0). \tag{3.76}$$

Similar conditions can be given for the setup in this chapter, namely one must have

$$-\log(p_{ij}) + T_{ij} + \lambda_j - \lambda_i \leq C \quad \text{(with equality if } Q_{ij} > 0), \tag{3.77}$$

for some constants λ_i, $i \in \mathcal{S}$.

Comment: Is it necessary to say more about the topic of Kuhn-Tucker conditions?

3.6 How to Compute the T_{ij} Values Efficiently

Remark 39 (Efficient Computation of T_{ij} Values) Although the definition of the T_{ij} values is quite complicated, they can indeed be computed quite efficiently. One can use the following steps to get the T_{ij} values with probability one as $N \to \infty$.

- Choose a large N.
- Randomly generates an input sequence and therefore a state sequence $\check{\mathbf{s}}$, and subsequently an output sequence $\check{\mathbf{y}}$. (With probability 1 as $N \to \infty$ each of these sequences are typical and together they are jointly typical.)
- For all $(i, j) \in \mathcal{A}$, $\ell \in \mathcal{I}_N$ compute $V_{\ell-1,\ell}(i, j|\check{\mathbf{y}})$, and for all $i \in \mathcal{S}$, $\ell \in \mathcal{I}'_N$ compute $V_\ell(i|\check{\mathbf{y}})$ using the BCJR (or forward-backward) algorithm [11].
- (First possibility) Compute

$$\check{T}_{ij}^{(N)}(\check{\mathbf{s}}, \check{\mathbf{y}}) = \overline{\overline{\check{T}}}_{ij}^{(N)}(\check{\mathbf{s}}, \check{\mathbf{y}}) - \overline{\check{T}}_{i}^{(N)}(\check{\mathbf{s}}, \check{\mathbf{y}}), \tag{3.78}$$

$$\begin{cases} \overline{\overline{\check{T}}}_{ij}^{(N)}(\check{\mathbf{s}}, \check{\mathbf{y}}) &= \frac{1}{N' Q_{ij}} \sum\limits_{\substack{\ell \in \mathcal{I}_N \\ \check{s}_{\ell-1}=i, \check{s}_\ell = j}} \log\left(V_{\ell-1,\ell}(i, j|\check{\mathbf{y}})\right), \\ \overline{\check{T}}_{i}^{(N)}(\check{\mathbf{s}}, \check{\mathbf{y}}) &= \frac{1}{N' \mu_i} \sum\limits_{\substack{\ell \in \mathcal{I}'_N \\ \check{s}_\ell = i}} \log\left(V_\ell(i|\check{\mathbf{y}})\right). \end{cases} \tag{3.79}$$

- (Second possibility) Computationally better (i.e. better accuracy for smaller N for the cases where some p_{ij}'s are low, i.e., the corresponding branches are visited rarely) give the computation rules (see also [1])

$$\check{T}_{ij}^{(N)}(\check{\mathbf{y}}) = \overline{\overline{\check{T}}}_{ij}^{(N)}(\check{\mathbf{y}}) - \overline{\check{T}}_{i}^{(N)}(\check{\mathbf{y}}), \tag{3.80}$$

$$\begin{cases} \overline{\overline{\check{T}}}_{ij}^{(N)}(\check{\mathbf{y}}) &= \frac{1}{N'} \sum\limits_{\ell \in \mathcal{I}_N} \frac{V_{\ell-1,\ell}(i,j|\check{\mathbf{y}})}{Q_{ij}} \log\left(V_{\ell-1,\ell}(i, j|\check{\mathbf{y}})\right), \\ \overline{\check{T}}_{i}^{(N)}(\check{\mathbf{y}}) &= \frac{1}{N'} \sum\limits_{\ell \in \mathcal{I}'_N} \frac{V_\ell(i|\check{\mathbf{y}})}{\mu_i} \log\left(V_\ell(i|\check{\mathbf{y}})\right). \end{cases} \tag{3.81}$$

  This second possibility is close in spirit to the approach taken in [12] to modify the usual procedure to get estimates of bit-error rates.

- With probability 1 as $N \to \infty$ the value $\check{T}_{ij}^{(N)}(\check{\mathbf{s}}, \check{\mathbf{y}})$ and $\check{T}_{ij}^{(N)}(\check{\mathbf{y}})$ will be equal to the desired T_{ij}.

Proof: See Sec. B.15. □

3.7 Symmetrization

The definition of T_{ij} in Def. 21 is slightly asymmetric, but it is possible to symmetrize the definition. Based on the same definitions of $\overline{\overline{T}}_{ij}$ and $\overline{T}_i$ as in Def. 21, we first propose a more general definition of T_{ij}. Instead of $T_{ij} = \overline{\overline{T}}_{ij} - \overline{T}_i$ we set (for any $\beta \in \mathbb{R}$)

$$T_{ij}^{[\beta]} \triangleq \overline{\overline{T}}_{ij} - \left(\beta\overline{T}_i + (1-\beta)\overline{T}_j\right). \tag{3.82}$$

After noting that

$$\sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{\overline{T}}_i = \sum_{i\in\mathcal{S}} \mu_i\tilde{\overline{T}}_i = \sum_{j\in\mathcal{S}} \mu_j\tilde{\overline{T}}_j = \sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{\overline{T}}_j, \tag{3.83}$$

we see that

$$\begin{aligned}
\sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{T}_{ij} &= \sum_{(i,j)\in\mathcal{A}} Q_{ij}(\tilde{\overline{\overline{T}}}_{ij} - \tilde{\overline{T}}_i) && (3.84)\\
&= \sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{\overline{\overline{T}}}_{ij} - \beta\sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{\overline{T}}_i - (1-\beta)\sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{\overline{T}}_i && (3.85)\\
&= \sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{\overline{\overline{T}}}_{ij} - \beta\sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{\overline{T}}_i - (1-\beta)\sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{\overline{T}}_j && (3.86)\\
&= \sum_{(i,j)\in\mathcal{A}} Q_{ij}\cdot\left[\tilde{\overline{\overline{T}}}_{ij} - \left(\beta\tilde{\overline{T}}_i + (1-\beta)\tilde{\overline{T}}_j\right)\right] = \sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{T}_{ij}^{[\beta]}. && (3.87)
\end{aligned}$$

So, with this new definition, expressions like $\sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{T}_{ij}$ can be replaced by $\sum_{(i,j)\in\mathcal{A}} Q_{ij}\tilde{T}_{ij}^{[\beta]}$. Simlilarly, we can show that $\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha}\tilde{T}_{ij} = \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha}\tilde{T}_{ij}^{[\beta]}$.

The choice made in Def. 21 corresponds to $\beta = 1$. But the choice $\beta = 1/2$ leads to the more symmetric definition

$$T_{ij}^{[1/2]} \triangleq \overline{\overline{T}}_{ij} - \frac{1}{2}\left(\overline{T}_i + \overline{T}_j\right). \tag{3.88}$$

If for some channel and some $\{Q_{ij}\}$ we have $\overline{\overline{T}}_{ij} = \overline{\overline{T}}_{ji}$ for all $(i,j) \in \mathcal{A}$ (this is a very special channel), then $\tilde{\mathbf{A}}$ is symmetric, i.e. all eigenvalues are real and the corresponding left and the right eigenvectors are equal.

3.8 Concavity of Various Functions

Lemma 40 *The functions $f_1^{(N)}(Q_{ij})$ and $f_1(Q_{ij})$ defined in Def. 18 are concave in $\{Q_{ij}\}$.*

Proof: See Sec. B.16. □

Conjecture 41 *We conjecture that $f_2(Q_{ij}, W), f_3(Q_{ij}, W), f_4(Q_{ij}, W), I(Q_{ij}, W)$ are concave in $\{Q_{ij}\}$.*

3.9 The Classical Blahut-Arimoto Algorithm as a Special Case of the Generalized Blahut-Arimoto Algorithm

A DMC can be brought into the form required to apply Alg. 37 by the following definitions. Assume that there are n input symbols and let p_j be the probability of sending symbol j. Let $\mathcal{S} = \{1, \ldots, n\}$, and there is a transition from state i to state j with probability p_{ij} and one sends input symbol j (so being in state j means that in the last transmission we sent symbol j). In this setting, $\tilde{T}_{ij}$ clearly only depends on j, so we set $\tilde{T}_{ij} = T_j$. We get the matrix

$$\tilde{\mathbf{A}} = \begin{pmatrix} \mathrm{e}^{\tilde{T}_{11}} & \cdots & \mathrm{e}^{\tilde{T}_{1n}} \\ \vdots & & \vdots \\ \mathrm{e}^{\tilde{T}_{n1}} & \cdots & \mathrm{e}^{\tilde{T}_{nn}} \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{\tilde{T}_{1}} & \cdots & \mathrm{e}^{\tilde{T}_{n}} \\ \vdots & & \vdots \\ \mathrm{e}^{\tilde{T}_{1}} & \cdots & \mathrm{e}^{\tilde{T}_{n}} \end{pmatrix}, \tag{3.89}$$

whose largest (real) eigenvalue is $\rho = \sum_j \mathrm{e}^{\tilde{T}_j}$ with right eigenvector $\mathbf{b}^\mathsf{T} = \mathbf{1}^\mathsf{T}$ and left eigenvector $\mathbf{c} = (\mathrm{e}^{\tilde{T}_1}, \ldots, \mathrm{e}^{\tilde{T}_n})$. Furthermore, $K = 1/\sum_j \mathrm{e}^{\tilde{T}_j}$ and $\mu_i = \mathrm{e}^{\tilde{T}_i}/\sum_{i'} \mathrm{e}^{\tilde{T}_{i'}}$. So the updated transition probabilities are

$$p^*_{ij} \triangleq \frac{b_j}{b_i} \cdot \frac{\mathrm{e}^{\tilde{T}_{ij}}}{\rho} = \frac{1}{1} \cdot \frac{\mathrm{e}^{\tilde{T}_j}}{\sum_{j'} \mathrm{e}^{\tilde{T}_{j'}}} = \frac{e^{\tilde{T}_j}}{\sum_{j'} \mathrm{e}^{\tilde{T}_{j'}}}, \tag{3.90}$$

which is independent of i; so we can set $p^*_j \triangleq p^*_{ij}$. If a capacity-achieving input distribution is achieved, the capacity is

$$C = \log(\rho) = \log\left(\sum_j \mathrm{e}^{\tilde{T}_j}\right). \tag{3.91}$$

3.10 Noiseless Channels with Memory as a Special Case of the Generalized Blahut-Arimoto Algorithm

Shannon [13] studied the case of noiseless channels with memory. In our setup, the capacity of such channels equals the normalized logarithm of the number of possible paths in the trellis representing the channel.

In that case $T_{ij} = 1$ if $(i,j) \in \mathcal{A}$. So in Alg. 37, the matrix $\tilde{\mathbf{A}}$ has an entry $\tilde{a}_{ij} = 0$ if $(i,j) \in \mathcal{A}$, and $\tilde{a}_{ij} = 0$ otherwise; i.e., $\tilde{\mathbf{A}}$ is a matrix consisting only of ones and zeros. This matrix is the adjacency matrix of the trellis. So, $\tilde{\mathbf{A}}$ in the general case can be considered as a *noisy adjacency matrix* [1].

We remark that in the noiseless case a Markov source having the same memory length as the channel is sufficient for achieving capacity (see also problem 13 of Chap. 4 in [14]). Judging from numerical results, this seems not to be the case in the noisy case: here the achievable information rates seem to increase as the Markov source memory length increases.

Chapter 4

Open Problems

4.1 Open Problems

The main open problems are the ones formulated in Conj. 41.

Appendix A

Proofs to Chapter 2

A.1 Proof of Lemma 5

For ease of notation we set $Q_x \triangleq Q(x)$. We will also need the general property

$$\sum_x Q_x^\alpha(\alpha) = \sum_x \frac{\mathrm{d}}{\mathrm{d}\alpha} Q_x(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} \sum_x Q_x(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} 1 = 0. \tag{A.1}$$

To prove concavity of the various functions, we express $Q_x \triangleq Q_x(\alpha)$ as a linear function in a single parameter α. If a function is concave in α for any such parametrization, then the function is concave in $Q(x)$. We have

$$Q_x^\alpha \triangleq Q_x^\alpha(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} Q_x(\alpha) \quad \text{are constants,} \tag{A.2}$$

$$Q_x^{\alpha\alpha} \triangleq \frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} Q_x(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} Q_x^\alpha(\alpha) = 0. \tag{A.3}$$

Subsequently, we will omit the argument α of $Q_x(\alpha)$. We start with proving that $H(X)$ is concave.

$$f_1(\alpha) \triangleq -\sum_x Q_x \log Q_x, \tag{A.4}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_1(\alpha) = -\sum_x Q_x^\alpha \log Q_x - \sum_x Q_x \frac{1}{Q_x} Q_x^\alpha = -\sum_x Q_x^\alpha \log Q_x, \tag{A.5}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} f_1(\alpha) = -\sum_x Q_x^{\alpha\alpha} \log Q_x - \sum_x Q_x^\alpha \frac{1}{Q_x} Q_x^\alpha = -\sum_x \frac{(Q_x^\alpha)^2}{Q_x} \leq 0. \tag{A.6}$$

We prove the concavity of $H(Y)$.

$$R(y) = (QW)(y) = \sum_x Q_x \sum_y W(y|x), \tag{A.7}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} R(y) = \sum_x Q_x^\alpha W(y|x). \tag{A.8}$$

$$f_2(\alpha) \stackrel{\triangle}{=} -\sum_y R(y)\log(R(y)) \tag{A.9}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_2(\alpha) = -\sum_y \left(\sum_x Q_x^\alpha W(y|x)\right)\log(R(y)) - \sum_y R(y)\frac{1}{R(y)}\left(\sum_x Q_x^\alpha W(y|x)\right) \tag{A.10}$$

$$= -\sum_y \left(\sum_x Q_x^\alpha W(y|x)\right)\log(R(y)) - \sum_x Q_x^\alpha \sum_y W(y|x) \tag{A.11}$$

$$= -\sum_y \left(\sum_x Q_x^\alpha W(y|x)\right)\log(R(y)) \tag{A.12}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} f_2(\alpha) = -\sum_y \left(\sum_x Q_x^{\alpha\alpha} W(y|x)\right)\log(R(y)) - \sum_y \left(\sum_x Q_x^\alpha W(y|x)\right)\frac{1}{R(y)}\left(\sum_x Q_x^\alpha W(y|x)\right) \tag{A.13}$$

$$= -\sum_y \frac{\left[\sum_x Q_x^\alpha W(y|x)\right]^2}{R(y)} \geq 0. \tag{A.14}$$

The concavity of the entropy of Y follows also from the concavity of X and the fact that $R(y)$ is linear function in $Q(x)$. We now prove the concavity of $H(Y|X)$.

$$f_3(\alpha) \stackrel{\triangle}{=} -\sum_x Q_x \sum_y W(y|x)\log(W(y|x)) \tag{A.15}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_3(\alpha) = -\sum_x Q_x^\alpha \sum_y W(y|x)\log(W(y|x)) \tag{A.16}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} f_3(\alpha) = -\sum_x Q_x^{\alpha\alpha} \sum_y W(y|x)\log(W(y|x)) = 0. \tag{A.17}$$

This result follows also from the fact that $H(Y|X)$ is linear in $Q(x)$. Therefore, $H(Y|X)$ is both concave and convex in $Q(.)$. From $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ we get $H(X|Y) = H(X) + H(Y|X) - H(Y)$ and so

$$f_4(\alpha) \stackrel{\triangle}{=} f_1(\alpha) + f_3(\alpha) - f_2(\alpha) \tag{A.18}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_4(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} f_1(\alpha) + \frac{\mathrm{d}}{\mathrm{d}\alpha} f_3(\alpha) - \frac{\mathrm{d}}{\mathrm{d}\alpha} f_2(\alpha), \tag{A.19}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} f_4(\alpha) = \frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} f_1(\alpha) + \frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} f_3(\alpha) - \frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} f_2(\alpha) \tag{A.20}$$

$$= -\sum_x \frac{(Q_x^\alpha)^2}{Q_x} + 0 + \sum_y \frac{\left[\sum_x Q_x^\alpha W(y|x)\right]^2}{R(y)}. \tag{A.21}$$

To proceed, we need the Cauchy-Schwarz inequality which says that

$$\left(\sum_x a_x b_x\right)^2 \leq \left(\sum_x a_x^2\right)\cdot\left(\sum_x b_x^2\right), \tag{A.22}$$

where equality holds if and only if $a_x = b_x$ for all x. With

$$a_x = Q_x^\alpha \sqrt{\frac{W(y|x)}{Q_x}}, \tag{A.23}$$

$$b_x = \sqrt{Q_x W(y|x)} \tag{A.24}$$

we get

$$\left[\sum_x Q_x^\alpha W(y|x)\right]^2 \leq \left[\sum_x (Q_x^\alpha)^2 \frac{W(y|x)}{Q_x}\right] \cdot \left[\sum_x Q_x W(y|x)\right] \tag{A.25}$$

$$= \left[\sum_x (Q_x^\alpha)^2 \frac{W(y|x)}{Q_x}\right] \cdot R(y)) \tag{A.26}$$

Using this side result we obtain

$$\frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} f_4(\alpha) = -\sum_x \frac{(Q_x^\alpha)^2}{Q_x} + \sum_y \frac{\left[\sum_x Q_x^\alpha W(y|x)\right]^2}{R(y)} \tag{A.27}$$

$$\leq -\sum_x \frac{(Q_x^\alpha)^2}{Q_x} + \sum_y \frac{\left[\sum_x (Q_x^\alpha)^2 \frac{W(y|x)}{Q_x}\right] \cdot R(y)}{R(y)} \tag{A.28}$$

$$= -\sum_x \frac{(Q_x^\alpha)^2}{Q_x} + \sum_x \frac{(Q_x^\alpha)^2}{Q_x} \sum_y W(y|x) \tag{A.29}$$

$$= -\sum_x \frac{(Q_x^\alpha)^2}{Q_x} + \sum_x \frac{(Q_x^\alpha)^2}{Q_x} = 0. \tag{A.30}$$

From $I(X;Y) = H(Y) - H(Y|X)$ we have

$$f_5(\alpha) \stackrel{\triangle}{=} f_2(\alpha) - f_3(\alpha), \tag{A.31}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_5(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} f_2(\alpha) - \frac{\mathrm{d}}{\mathrm{d}\alpha} f_3(\alpha), \tag{A.32}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} f_5(\alpha) = \frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} f_2(\alpha) - \frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} f_3(\alpha) \tag{A.33}$$

$$= -\sum_y \frac{\left[\sum_x Q_x^\alpha W(y|x)\right]^2}{R(y)} - 0 \leq 0. \tag{A.34}$$

A.2 Proof of Lemma 8

The first statement is clear from Defs. 2 and 7. The second statement follows from

$$
\begin{aligned}
& I(Q,W) - \Psi(\tilde{Q},Q,W) && \text{(A.35)}\\
&= \Psi(Q,Q,W) - \Psi(\tilde{Q},Q,W) && \text{(A.36)}\\
&= \sum_{x,y} Q(x)W(y|x)\log\left(\frac{V(x|y))}{Q(x)}\right) - \sum_{x,y} Q(x)W(y|x)\log\left(\frac{\tilde{V}(x|y)}{Q(x)}\right) && \text{(A.37)}\\
&= \sum_{x,y} Q(x)W(y|x)\log\left(\frac{V(x|y)}{\tilde{V}(x|y)}\right) = \sum_y R(y)\sum_x V(x|y)\log\left(\frac{V(x|y)}{\tilde{V}(x|y)}\right) && \text{(A.38)}\\
&= \sum_y R(y) D\big(V(\cdot|y)\,||\,\tilde{V}(\cdot|y)\big) \geq \sum_y R(y)\cdot 0 = 0, && \text{(A.39)}
\end{aligned}
$$

where the inequality follows from the fact that relative entropies are non-negative. The first inequality of the third statement in Lemma 8 follows from the the fact that we can choose $Q = \tilde{Q}$, whereupon $\Psi(\tilde{Q},Q,W) = \Psi(\tilde{Q},\tilde{Q},W) = I(\tilde{Q},W)$ so that $\Psi(\tilde{Q},Q,W) \geq I(\tilde{Q},W)$ holds, whereas the second inequality of the third statement of Lemma 8 follows from the second statement.

A.3 Proof of Remark 9

Deriving

$$
\begin{aligned}
f_4(Q,W) &= -\sum_x Q(x)\sum_y W(y|x)\log(V(x|y)) && \text{(A.40)}\\
&= -\sum_x Q(x)\sum_y W(y|x)\log\left(\frac{Q(x)W(y|x)}{\sum_{x'} Q(x')W(y|x')}\right) && \text{(A.41)}
\end{aligned}
$$

with respect to $Q(x)$ leads to

$$
\begin{aligned}
\frac{\partial}{\partial Q(x)} f_4(Q,W) &= -\sum_y W(y|x)\log(V(x|y)) - Q(x)\sum_y W(y|x)\frac{1}{Q(x)} && \text{(A.42)}\\
&\quad + \sum_{x''} Q(x'')\sum_y W(y|x'')\frac{1}{\sum_{x'} Q(x')W(y|x')} W(y|x) && \text{(A.43)}\\
&= -\sum_y W(y|x)\log(V(x|y)) - 1 + 1 = -\sum_y W(y|x)\log(V(x|y)). && \text{(A.44)}
\end{aligned}
$$

Evaluated at $Q = \tilde{Q}$ this gives the desired result. We could have alternatively taken the results from Sec. A.1.

A.4 Proof of Theorem 11

We first prove the first statement. Fix some non-negative integer r. From (2.25) we know that $\Psi\big(Q^{\langle r\rangle}, Q^{\langle r\rangle}, W\big) = I(Q^{\langle r\rangle}, W)$. Taking the Q^* that maximizes $\Psi\big(Q^{\langle r\rangle}, Q, W\big)$ we must have

$\Psi\big(Q^{\langle r\rangle}, Q^*, W\big) \geq I(Q^{\langle r\rangle}, W)$. But from (2.26) we know also that $\Psi\big(Q^{\langle r\rangle}, Q^*, W\big) \leq I(Q^*, W)$. Combining these two results we have

$$I(Q^{\langle r\rangle}, W) \leq \Psi\big(Q^{\langle r\rangle}, Q^*, W\big) \leq I(Q^*, W). \tag{A.45}$$

By showing that $Q^{(r+1)} = Q^*$ we finish the proof of the first statement. For maximizing $\Psi(Q^{\langle r\rangle}, Q, W)$ over Q under the constraint $\sum_x Q(x) = 1$ we use Lagrange multipliers. Solving

$$0 \stackrel{!}{=} \frac{\partial}{\partial Q(x)} \left(\Psi\big(Q^{\langle r\rangle}, Q, W\big) + \lambda \sum_{x'} Q(x') \right) \Bigg|_{Q=Q^*} \tag{A.46}$$

$$= -\log(Q^*(x)) - Q^*(x)/Q^*(x) + T^{\langle r\rangle}(x) + \lambda \tag{A.47}$$

we get for each x

$$Q^*(x) = \frac{\mathrm{e}^{T^{\langle r\rangle}(x)}}{\sum_{x'} \mathrm{e}^{T^{\langle r\rangle}(x')}}. \tag{A.48}$$

We now prove the second statement, namely that $Q^{\langle r\rangle}$ converges to a capacity-achieving input distribution for $r \to \infty$. Let $Q^{\langle 0\rangle}(x)$ be some initial (freely chosen) input distribution. We use the definitions

$$V^{\langle r\rangle}(x|y) \triangleq \frac{W(y|x)Q^{\langle r\rangle}(x)}{(QW)(y)}, \tag{A.49}$$

$$T^{\langle r\rangle}(x) \triangleq \sum_y W(y|x) \log\left(V^{\langle r\rangle}(x|y)\right), \tag{A.50}$$

$$Q^{(r+1)}(x) \triangleq \frac{\exp\left(T^{\langle r\rangle}(x)\right)}{\sum_{x'} \exp\left(T^{\langle r\rangle}(x')\right)}, \tag{A.51}$$

$$C^{\langle r,r\rangle} \triangleq I\left(Q^{\langle r\rangle}, W\right) = \Psi\left(Q^{\langle r\rangle}, Q^{\langle r\rangle}, W\right) = \sum_x Q^{\langle r\rangle}(x) \sum_y W(y|x) \log\left(\frac{V^{\langle r\rangle}(x|y)}{Q^{\langle r\rangle}(x)}\right), \tag{A.52}$$

$$C^{\langle r,r+1\rangle} \triangleq \Psi\left(Q^{\langle r\rangle}, Q^{(r+1)}, W\right) = \sum_x Q^{(r+1)}(x) \sum_y W(y|x) \log\left(\frac{V^{\langle r\rangle}(x|y)}{Q^{(r+1)}(x)}\right). \tag{A.53}$$

From Lemma 8,

$$C^{\langle 0,0\rangle} \leq C^{\langle 0,1\rangle} \leq \cdots \leq C^{\langle r,r\rangle} \leq C^{\langle r,r+1\rangle} \leq C^{\langle r+1,r+1\rangle} \leq \cdots \leq C. \tag{A.54}$$

Note that $C^{\langle r,r\rangle}$ is the information rate when the input has the distribution $Q^{\langle r\rangle}(\cdot)$. We observe that for any x

$$\begin{aligned}
&\sum_y W(y|x)\log\left(\frac{V^{\langle r\rangle}(x|y)}{Q^{(r+1)}(x)}\right) && \text{(A.55)}\\
&= \left(\sum_y W(y|x)\log\left(V^{\langle r\rangle}(x|y)\right)\right) - \log\left(Q^{(r+1)}(x)\right)\underbrace{\sum_y W(y|x)}_{=1} && \text{(A.56)}\\
&= \underbrace{\left(\sum_y W(y|x)\log\left(V^{\langle r\rangle}(x|y)\right)\right) - T^{\langle r\rangle}(x)}_{=0} + \log\left(\sum_{x'}\exp(T^{\langle r\rangle}(x'))\right) && \text{(A.57)}\\
&= \log\left(\sum_{x'}\exp(T^{\langle r\rangle}(x'))\right), && \text{(A.58)}
\end{aligned}$$

which is independent of x. So, using this result we get

$$\begin{aligned}
C^{\langle r,r+1\rangle} &= \sum_x Q^{(r+1)}(x)\sum_y W(y|x)\log\left(\frac{V^{\langle r\rangle}(x|y)}{Q^{(r+1)}(x)}\right) && \text{(A.59)}\\
&= \log\left(\sum_{x'}\exp(T^{\langle r\rangle}(x'))\right). && \text{(A.60)}
\end{aligned}$$

Moreover we have

$$\begin{aligned}
\exp\left(T^{\langle r\rangle}(x)\right) &= \exp\left(\sum_y W(y|x)\log\left(\frac{W(y|x)Q^{\langle r\rangle}(x)}{(Q^{\langle r\rangle}W)(y)}\right)\right) && \text{(A.61)}\\
&= \exp\left(\sum_y W(y|x)\log\left(\frac{W(y|x)}{(Q^{\langle r\rangle}W)(y)}\right) + \log Q^{\langle r\rangle}(x)\sum_y W(y|x)\right) && \text{(A.62)}\\
&= Q^{\langle r\rangle}(x)\cdot\exp\left(\sum_y W(y|x)\log\left(\frac{W(y|x)}{(Q^{\langle r\rangle}W)(y)}\right)\right) && \text{(A.63)}
\end{aligned}$$

Combining (A.60) and (A.63), the update rule reads

$$\begin{aligned}
Q^{(r+1)}(x) &= \frac{\exp\left(T^{\langle r\rangle}(x)\right)}{\sum_{x'}\exp\left(T^{\langle r\rangle}(x')\right)} && \text{(A.64)}\\
&= \frac{Q^{\langle r\rangle}(x)\cdot\exp\left(\sum_y W(y|x)\log\left(\frac{W(y|x)}{(Q^{\langle r\rangle}W)(y)}\right)\right)}{\exp\left(C^{\langle r,r+1\rangle}\right)} && \text{(A.65)}
\end{aligned}$$

Let $C = C(W)$ be the capacity and $Q^*(\cdot)$ be a capacity-achieving input distribution. Using

Equation (A.65) we observe that

$$\sum_x Q^*(x)\log\left(\frac{Q^{(r+1)}(x)}{Q^{\langle r\rangle}(x)}\right) \tag{A.66}$$

$$= -C^{\langle r,r+1\rangle} + \sum_x Q^*(x)\sum_y W(y|x)\log\left(\frac{W(y|x)}{(Q^{\langle r\rangle}W)(y)}\right) \tag{A.67}$$

$$= -C^{\langle r,r+1\rangle} + \sum_x\sum_y Q^*(x)W(y|x)\log\left(\frac{W(y|x)}{(Q^*W)(y)}\cdot\frac{(Q^*W)(y)}{(Q^{\langle r\rangle}W)(y)}\right) \tag{A.68}$$

$$= -C^{\langle r,r+1\rangle} + \sum_x\sum_y Q^*(x)W(y|x)\log\left(\frac{W(y|x)}{(Q^*W)(y)}\right) \tag{A.69}$$

$$+ \sum_y (Q^*W)(y)\log\left(\frac{(Q^*W)(y)}{(Q^{\langle r\rangle}W)(y)}\right) \tag{A.70}$$

$$= -C^{\langle r,r+1\rangle} + C + D\left((Q^*W)(y)\middle\|(Q^{\langle r\rangle}W)(y)\right) \tag{A.71}$$

$$\geq -C^{\langle r,r+1\rangle} + C + 0 = C - C^{\langle r,r+1\rangle}. \tag{A.72}$$

Therefore

$$C - C^{\langle r,r+1\rangle} \leq \sum_x Q^*(x)\log\left(\frac{Q^{(r+1)}(x)}{Q^{\langle r\rangle}(x)}\right), \tag{A.73}$$

and summing over r from 0 to some N

$$\sum_{r=0}^{N}\left(C - C^{\langle r,r+1\rangle}\right) \leq \sum_{r=0}^{N}\sum_x Q^*(x)\log\left(\frac{Q^{(r+1)}(x)}{Q^{\langle r\rangle}(x)}\right) \tag{A.74}$$

$$= \sum_x Q^*(x)\log\left(\frac{Q^{(N+1)}(x)}{Q^{\langle 0\rangle}(x)}\right) \tag{A.75}$$

$$= \sum_x Q^*(x)\log\left(\frac{Q^*(x)}{Q^{\langle 0\rangle}(x)}\cdot\frac{Q^{(N+1)}(x)}{Q^*(x)}\right) \tag{A.76}$$

$$= D\big(Q^*(\cdot)||Q^{\langle 0\rangle}(\cdot)\big) - D\big(Q^*(\cdot)||Q^{(N+1)}(\cdot)\big) \tag{A.77}$$

$$\leq D\big(Q^*(\cdot)||Q^{\langle 0\rangle}(\cdot)\big), \tag{A.78}$$

where the right hand side is independent of N. A sufficient condition for $D\big(Q^*(\cdot)||Q^{\langle 0\rangle}(\cdot)\big)$ to be finite is that $Q^{\langle 0\rangle}(x)$ is larger than zero for each x. As $C - C^{\langle r,r+1\rangle}$ is non-negative for all r, we must have

$$C = \lim_{r\to\infty} C^{\langle r,r+1\rangle} = \lim_{r\to\infty} C^{\langle r,r\rangle}. \tag{A.79}$$

Note that we also have

$$\lim_{r\to\infty} D\left(\sum_x W(.|x)Q^*(x)\middle\|\sum_x W(.|x)Q^{\langle r\rangle}(x)\right) = 0, \tag{A.80}$$

i.e., the "capacity-achieving output distribution" is unique also when the capacity-achieving input distribution is non-unique. But if there if an unique $Q^*(\cdot)$ that achieves capacity, then $Q^{\langle r\rangle}(\cdot)$ converges to it.

This proof was essentially motivated by the ones given in the papers by Arimoto [8] and by O'Sullivan [9]. The proof idea in the paper by Blahut [7] is essentially along the same lines.

A.5 Proof of Remark 12

Let C be the capacity of the DMC with channel law $W(y|x)$ and let $Q(.)$ be any input pmf. Then,

$$
\begin{aligned}
& C - \min_x \left[T(x) - \log(Q(x))\right] && \text{(A.81)}\\
&= C - \min_x \left[\sum_y W(y|x)\log(V(x|y)) - \log(Q(x))\right] && \text{(A.82)}\\
&= C - \min_x \left[\sum_y W(y|x)\log\left(\frac{V(x|y)}{Q(x)}\right)\right] && \text{(A.83)}\\
&= C - \min_x \left[\sum_y W(y|x)\log\left(\frac{W(y|x)}{(QW)(y)}\right)\right] && \text{(A.84)}\\
&\overset{(*)}{\geq} C - \sum_x Q(x)\sum_y W(y|x)\log\left(\frac{W(y|x)}{(QW)(y)}\right) && \text{(A.85)}\\
&= C - I(Q,W) && \text{(A.86)}\\
&\overset{(**)}{\geq} 0, && \text{(A.87)}
\end{aligned}
$$

where inequality $(*)$ follows from the fact that a weighted sum (where the non-negative weights sum to one) is never smaller than its smallest term and inequality $(**)$ follows from the fact that capacity is always at least as large as any information rate. This result gives the first two inequalities in the Lemma.

Let $Q^*(\cdot)$ be a capacity-achieving input distribution and $Q(\cdot)$ any input distribution.

Then,

$$\max_x \left[T(x) - \log(Q(x))\right] - C \tag{A.88}$$

$$= \max_x \left[\sum_y W(y|x)\log(V(x|y)) - \log(Q(x))\right] - C \tag{A.89}$$

$$= \max_x \left[\sum_y W(y|x)\log\left(\frac{V(x|y)}{Q(x)}\right)\right] - C \tag{A.90}$$

$$= \max_x \left[\sum_y W(y|x)\log\left(\frac{W(y|x)}{(QW)(y)}\right)\right] - C \tag{A.91}$$

$$\overset{(*)}{\geq} \sum_x Q^*(x)\sum_y W(y|x)\log\left(\frac{W(y|x)}{(QW)(y)}\right) - C \tag{A.92}$$

$$= \sum_x Q^*(x)\sum_y W(y|x)\log\left(\frac{W(y|x)}{(QW)(y)}\right) - \sum_x Q^*(x)\sum_y W(y|x)\log\left(\frac{W(y|x)}{(Q^*W)(y)}\right) \tag{A.93}$$

$$= \sum_y Q^*W(y)\log\left(\frac{(Q^*W)(y)}{(QW)(y)}\right) \tag{A.94}$$

$$\overset{(**)}{\geq} 0, \tag{A.95}$$

where inequality $(*)$ follows from the fact that a weighted sum (where the non-negative weights sum to one) is never larger than its larges term and inequality $(**)$ follows from the fact that relative entropy is always at least zero. This gives the last inequality in the lemma.

Appendix B

Proofs to Chapter 3

B.1 Some Auxiliary Lemmas

Lemma 42 (Markov Property of A Posteriori PMFs) *Using the hidden Markov model structure of $Q(\mathbf{s})W(\mathbf{y}|\mathbf{s})$, we have for each $\ell \in \{-N, \ldots, N\}$*

$$V(s_\ell|\mathbf{s}_{-N}^{\ell-1}, \mathbf{y}) = V(s_\ell|s_{\ell-1}, \mathbf{y}) = V(s_\ell|s_{\ell-1}, \mathbf{y}_\ell^N), \tag{B.1}$$
$$V(s_\ell|\mathbf{s}_{\ell+1}^{N}, \mathbf{y}) = V(s_\ell|s_{\ell+1}, \mathbf{y}) = V(s_\ell|s_{\ell+1}, \mathbf{y}_{-N+1}^{\ell+1}), \tag{B.2}$$

i.e., given $\mathbf{y}$, $V(\mathbf{s}|\mathbf{y})$ is a Markov probability distribution in $\mathbf{s}$, i.e. Using these properties, we can rewrite $V(\mathbf{s}|\mathbf{y})$ in two useful different ways as products, namely,

$$\begin{aligned} V(\mathbf{s}|\mathbf{y}) &= V(s_{\ell-1}, s_\ell|\mathbf{y}) \cdot V(\mathbf{s}_{\ell+1}^N|s_\ell, \mathbf{y}) \cdot V(\mathbf{s}_{-N}^{\ell-2}|s_{\ell-1}, \mathbf{y}) && \text{(B.3)}\\ &= V(s_{\ell-1}, s_\ell|\mathbf{y}) \cdot V(\mathbf{s}_{\ell+1}^N|s_\ell, \mathbf{y}_{\ell+1}^N) \cdot V(\mathbf{s}_{-N}^{\ell-2}|s_{\ell-1}, \mathbf{y}_{-N+1}^{\ell-1}), && \text{(B.4)}\\ V(\mathbf{s}|\mathbf{y}) &= V(s_\ell|\mathbf{y}) \cdot V(\mathbf{s}_{\ell+1}^N|s_\ell, \mathbf{y}) \cdot V(\mathbf{s}_{-N}^{\ell-1}|s_\ell, \mathbf{y}) && \text{(B.5)}\\ &= V(s_\ell|\mathbf{y}) \cdot V(\mathbf{s}_{\ell+1}^N|s_\ell, \mathbf{y}_{\ell+1}^N) \cdot V(\mathbf{s}_{-N}^{\ell-1}|s_\ell, \mathbf{y}_{-N+1}^{\ell}). && \text{(B.6)} \end{aligned}$$

Proof: See Sec. B.17. □

Remark 43 (On the Markov Property of A Posteriori PMFs) Considering the corresponding (Forney-style) factor graph (normal graph) [15, 16] of the hidden Markov model (see Fig. B.1), the statements in Lemma 42 are rather straightforward.

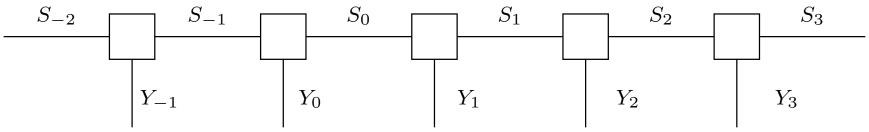

Figure B.1: Forney-style factor graph (normal graph) representing a hidden Markov model.

Lemma 44 (Some Derivatives) *With the setup as in Def. 14 we have*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}Q(\mathbf{s}) = \left(\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell\in\mathcal{I}_N\\ s_{\ell-1}=i,s_\ell=j}} \frac{Q(\mathbf{s})}{Q_{ij}}\right) - \left(\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell\in\mathcal{I}'_N\\ s_\ell=i}} \frac{Q(\mathbf{s})}{\mu_i}\right), \tag{B.7}$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_{\mathbf{s}} Q(\mathbf{s})\log Q(\mathbf{s}) = \sum_{\mathbf{s}}\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}Q(\mathbf{s})\right)\log Q(\mathbf{s}). \tag{B.8}$$

Proof: See Sec. B.18. □

B.2 Proof of Lemma 22

Using the definition of $Q(\mathbf{s})$ in (3.7) we obtain the first statement.

$$\frac{1}{N'}\sum_{\mathbf{s}} Q(\mathbf{s})\log Q(\mathbf{s}) \tag{B.9}$$

$$= \frac{1}{N'}\sum_{\mathbf{s}} Q(\mathbf{s})\log\left(\mu_{s_{-N}}\prod_{\ell\in\mathcal{I}_N} p_{s_{\ell-1},s_\ell}\right) \tag{B.10}$$

$$= \frac{1}{N'}\sum_{\mathbf{s}} Q(\mathbf{s})\log\left(\mu_{s_{-N}}\right) + \frac{1}{N'}\sum_{\ell\in\mathcal{I}_N}\sum_{\mathbf{s}} Q(\mathbf{s})\log\left(p_{s_{\ell-1},s_\ell}\right) \tag{B.11}$$

$$= \frac{1}{N'}\sum_{s_{-N}} Q(s_{-N})\log\left(\mu_{s_{-N}}\right) + \frac{1}{N'}\sum_{\ell\in\mathcal{I}_N}\sum_{s_{\ell-1},s_\ell\in\mathcal{A}} Q(s_{\ell-1},s_\ell)\log\left(p_{s_{\ell-1},s_\ell}\right) \tag{B.12}$$

$$= \frac{1}{N'}\sum_{i\in\mathcal{S}} \mu_i\log\left(\mu_i\right) + \frac{1}{N'}\sum_{\ell\in\mathcal{I}_N}\sum_{(i,j)\in\mathcal{A}} Q_{ij}\log\left(p_{ij}\right) \tag{B.13}$$

$$= \frac{1}{N'}\sum_{i\in\mathcal{S}} \mu_i\log\left(\mu_i\right) + \sum_{(i,j)\in\mathcal{A}} Q_{ij}\log\left(p_{ij}\right), \tag{B.14}$$

$$= \frac{1}{N'}\sum_{(i,j)\in\mathcal{A}} Q_{ij}\log\left(\mu_i\right) + \sum_{(i,j)\in\mathcal{A}} Q_{ij}\log\left(p_{ij}\right). \tag{B.15}$$

This essentially follows also from $H(X_{-N},\ldots,X_N) = H(X_{-N})+\sum_{\ell\in\mathcal{I}_N} H(X_\ell|X_{-N},\ldots,H_{\ell-1}) = H(X_{-N}) + \sum_{\ell\in\mathcal{I}_N} H(X_\ell|H_{\ell-1})$.

B.3 Proof of Lemma 23

We want to derive

$$f_1^{(N)}(\{Q_{ij}\}) = -\sum_{(i,j)\in\mathcal{A}} Q_{ij}\log\left(p_{ij}\right) - \frac{1}{N'}\sum_{(i,j)\in\mathcal{A}} Q_{ij}\log\left(\mu_i\right) \tag{B.16}$$

with respect to α.

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\alpha} f_1^{(N)}(\alpha) && \text{(B.17)}\\
&= -\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \log\left(p_{ij}\right) - \sum_{(i,j)\in\mathcal{A}} Q_{ij} \frac{1}{p_{ij}} p_{ij}^{\alpha} - \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \log(\mu_i) - \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij} \frac{1}{\mu_i} \mu_i^{\alpha} && \text{(B.18)}\\
&= -\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \log\left(p_{ij}\right) - \sum_{i\in\mathcal{S}} \mu_i \sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij}^{\alpha} - \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \log(\mu_i) - \frac{1}{N'} \sum_{i\in\mathcal{S}} \mu_i^{\alpha} \underbrace{\sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij}}_{=1} && \text{(B.19)}\\
&= -\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \log\left(p_{ij}\right) - \sum_{i\in\mathcal{S}} \mu_i \frac{\mathrm{d}}{\mathrm{d}\alpha} \underbrace{\sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij}}_{=1} - \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \log(\mu_i) - \frac{1}{N'} \frac{\mathrm{d}}{\mathrm{d}\alpha} \sum_{i\in\mathcal{S}} \mu_i && \text{(B.20)}\\
&= -\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \log\left(p_{ij}\right) - \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \log(\mu_i). && \text{(B.21)}
\end{aligned}$$

B.4 Proof of Lemma 24

We would like to rewrite the function $f_4^{(N)}(Q_{ij}, W)$ given in Def. 18, i.e.,[1]

$$-f_4^{(N)}(Q_{ij}, W) = \frac{1}{N'} \sum_{\mathbf{s}} Q(\mathbf{s}) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(\mathbf{s}|\mathbf{y})\right), \tag{B.22}$$

using the T_{ij}'s as given in Def. 21.

$$\begin{aligned}
V(\mathbf{s}|\mathbf{y}) &= V(s_{-N}|\mathbf{y}) \cdot \prod_{\ell\in\mathcal{I}_N} V(s_\ell|\mathbf{s}_{-N}^{\ell-1}, \mathbf{y}) && \text{(B.23)}\\
&\overset{(*)}{=} V(s_{-N}|\mathbf{y}) \cdot \prod_{\ell\in\mathcal{I}_N} V(s_\ell|s_{\ell-1}, \mathbf{y}) && \text{(B.24)}\\
&= V(s_{-N}|\mathbf{y}) \cdot \prod_{\ell\in\mathcal{I}_N} \frac{V(s_{\ell-1}, s_\ell|\mathbf{y})}{V(s_{\ell-1}|\mathbf{y})} = \left(\prod_{\ell\in\mathcal{I}_N} V(s_{\ell-1}, s_\ell|\mathbf{y})\right) \cdot \left(\prod_{\ell\in\mathcal{I}'_N} V(s_{\ell-1}|\mathbf{y})\right)^{-1}, && \text{(B.25)}
\end{aligned}$$

[1]For convenience reasons, we consider $-f_4^{(N)}(Q_{ij}, W)$ instead of $f_4^{(N)}(Q_{ij}, W)$.

where equality $(*)$ follows from Lemma 42, and so

$$
\begin{aligned}
&-f_4^{(N)}(Q_{ij},W) && \text{(B.26)}\\
&= \frac{1}{N'}\sum_{\ell\in\mathcal{I}_N}\sum_{\mathbf{s}} Q(\mathbf{s})\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(V(s_{\ell-1},s_\ell|\mathbf{y})\right) && \text{(B.27)}\\
&\quad -\frac{1}{N'}\sum_{\ell\in\mathcal{I}'_N}\sum_{\mathbf{s}} Q(\mathbf{s})\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(V(s_\ell|\mathbf{y})\right) && \text{(B.28)}\\
&= \frac{1}{N'}\sum_{\ell\in\mathcal{I}_N}\sum_{(i,j)\in\mathcal{A}} Q_{ij}\sum_{\substack{\mathbf{s}\\ s_{\ell-1}=i,s_\ell=j}} Q(\mathbf{s}_{-N}^{\ell-2},\mathbf{s}_{\ell+1}^N|s_{\ell-1},s_\ell)\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(V(s_{\ell-1},s_\ell|\mathbf{y})\right) && \text{(B.29)}\\
&\quad -\frac{1}{N'}\sum_{\ell\in\mathcal{I}'_N}\sum_{i\in\mathcal{S}}\mu_i\sum_{\substack{\mathbf{s}\\ s_\ell=i}} Q(\mathbf{s}_{-N}^{\ell-1},\mathbf{s}_{\ell+1}^N|s_\ell)\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(V(s_\ell|\mathbf{y})\right) && \text{(B.30)}\\
&= \sum_{(i,j)\in\mathcal{A}} Q_{ij}\frac{1}{N'}\sum_{\ell\in\mathcal{I}_N}\sum_{\substack{\mathbf{s}\\ s_{\ell-1}=i,s_\ell=j}} Q(\mathbf{s}_{-N}^{\ell-2},\mathbf{s}_{\ell+1}^N|s_{\ell-1},s_\ell)\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(V(s_{\ell-1},s_\ell|\mathbf{y})\right) && \text{(B.31)}\\
&\quad -\sum_{i\in\mathcal{S}}\mu_i\frac{1}{N'}\sum_{\ell\in\mathcal{I}'_N}\sum_{\substack{\mathbf{s}\\ s_\ell=i}} Q(\mathbf{s}_{-N}^{\ell-1},\mathbf{s}_{\ell+1}^N|s_\ell)\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(V(s_\ell|\mathbf{y})\right) && \text{(B.32)}\\
&= \sum_{(i,j)\in\mathcal{A}} Q_{ij}\overline{\overline{T}}_{ij}^{(N)} - \sum_{i\in\mathcal{S}}\mu_i\overline{T}_i^{(N)}. && \text{(B.33)}
\end{aligned}
$$

Using $\sum_{i\in\mathcal{S}}\mu_i = \sum_{(i,j)\in\mathcal{A}} Q_{ij}$ this can be cast the into the form

$$
-f_4^{(N)}(Q_{ij},W) = \sum_{(i,j)\in\mathcal{A}} Q_{ij}\overline{\overline{T}}_{ij}^{(N)} - \sum_{(i,j)\in\mathcal{A}} Q_{ij}\overline{T}_i^{(N)} = \sum_{(i,j)\in\mathcal{A}} Q_{ij}\cdot T_{ij}^{(N)}. \tag{B.34}
$$

B.5 Proof of Lemma 25

If the reader has not yet looked at Lemmas 42 and 44, we recommend to to so, because their results will be used here.

One must be very careful when deriving $V(\mathbf{s}|\mathbf{y})$ with respect to α, as $V(\mathbf{s}|\mathbf{y})$ depends on $Q(\mathbf{s})$, which depends on Q_{ij}, which depend on α, see also (3.7)-(3.10).

Deriving

$$
\begin{aligned}
-f_4^{(N)}(\alpha,W) &= \frac{1}{N'}\sum_{\mathbf{s}} Q(\mathbf{s})\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(V(\mathbf{s}|\mathbf{y})\right) && \text{(B.35)}\\
&= \frac{1}{N'}\sum_{\mathbf{s}} Q(\mathbf{s})\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})\log\left(\frac{Q(\mathbf{s})W(\mathbf{y}|\mathbf{s})}{R(\mathbf{y})}\right) && \text{(B.36)}
\end{aligned}
$$

with respect to α yields (in the following, we will use the abbreviation $J \triangleq -\frac{\mathrm{d}}{\mathrm{d}\alpha} f_4^{(N)}(\alpha, W)$)

$$
\begin{aligned}
J &= \frac{1}{N'} \sum_{\mathbf{s}} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log \left(\frac{Q(\mathbf{s}) W(\mathbf{y}|\mathbf{s})}{R(\mathbf{y})} \right) && \text{(B.37)} \\
&\quad + \frac{1}{N'} \sum_{\mathbf{s}} Q(\mathbf{s}) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \frac{1}{Q(\mathbf{s})} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) && \text{(B.38)} \\
&\quad - \frac{1}{N'} \sum_{\mathbf{s}} Q(\mathbf{s}) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \frac{1}{R(\mathbf{y})} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} R(\mathbf{y}) \right) && \text{(B.39)} \\
&= \frac{1}{N'} \sum_{\mathbf{s}} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log \left(V(\mathbf{s}|\mathbf{y}) \right) && \text{(B.40)} \\
&\quad + \frac{1}{N'} \sum_{\mathbf{s}} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) \underbrace{\sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s})}_{=1} - \frac{1}{N'} \sum_{\mathbf{y}} \frac{R(\mathbf{y})}{R(\mathbf{y})} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} R(\mathbf{y}) \right) && \text{(B.41)} \\
&= \frac{1}{N'} \sum_{\mathbf{s}} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log \left(V(\mathbf{s}|\mathbf{y}) \right) && \text{(B.42)} \\
&\quad + \frac{1}{N'} \frac{\mathrm{d}}{\mathrm{d}\alpha} \underbrace{\sum_{\mathbf{s}} Q(\mathbf{s})}_{=1} - \frac{1}{N'} \frac{\mathrm{d}}{\mathrm{d}\alpha} \underbrace{\sum_{\mathbf{y}} R(\mathbf{y})}_{=1} && \text{(B.43)} \\
&= \frac{1}{N'} \sum_{\mathbf{s}} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log \left(V(\mathbf{s}|\mathbf{y}) \right) && \text{(B.44)} \\
&\overset{(*)}{=} \frac{1}{N'} \sum_{\mathbf{s}} \left[\left(\sum_{(i,j) \in \mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell \in \mathcal{I}_N \\ s_{\ell-1}=i, s_\ell = j}} \frac{Q(\mathbf{s})}{Q_{ij}} \right) - \left(\sum_{(i,j) \in \mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell \in \mathcal{I}'_N \\ s_\ell = i}} \frac{Q(\mathbf{s})}{\mu_i} \right) \right] && \text{(B.45)} \\
&\quad \times \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log \left(V(\mathbf{s}|\mathbf{y}) \right) && \text{(B.46)}
\end{aligned}
$$

where equality $(*)$ follows from (B.7). We continue to evaluated J,

$$J \overset{(*)}{=} \frac{1}{N'} \sum_{\mathbf{s}} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell\in\mathcal{I}_N \\ s_{\ell-1}=i, s_\ell=j}} \frac{Q(\mathbf{s})}{Q_{ij}} \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_{\ell-1}, s_\ell|\mathbf{y})\right) \tag{B.47}$$

$$+ \frac{1}{N'} \sum_{\mathbf{s}} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell\in\mathcal{I}_N \\ s_{\ell-1}=i, s_\ell=j}} \frac{Q(\mathbf{s})}{Q_{ij}} \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(\mathbf{s}_{\ell+1}^{N}|s_\ell, \mathbf{y}_{\ell+1}^{N})\right) \tag{B.48}$$

$$+ \frac{1}{N'} \sum_{\mathbf{s}} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell\in\mathcal{I}_N \\ s_{\ell-1}=i, s_\ell=j}} \frac{Q(\mathbf{s})}{Q_{ij}} \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(\mathbf{s}_{-N}^{\ell-2}|s_{\ell-1}, \mathbf{y}_{-N+1}^{\ell-1})\right) \tag{B.49}$$

$$- \frac{1}{N'} \sum_{\mathbf{s}} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell\in\mathcal{I}'_N \\ s_\ell=i}} \frac{Q(\mathbf{s})}{\mu_i} \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_\ell|\mathbf{y})\right) \tag{B.50}$$

$$- \frac{1}{N'} \sum_{\mathbf{s}} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell\in\mathcal{I}'_N \\ s_\ell=i}} \frac{Q(\mathbf{s})}{\mu_i} \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(\mathbf{s}_{\ell+1}^{N}|s_\ell, \mathbf{y}_{\ell+1}^{N})\right) \tag{B.51}$$

$$- \frac{1}{N'} \sum_{\mathbf{s}} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell\in\mathcal{I}'_N \\ s_\ell=i}} \frac{Q(\mathbf{s})}{\mu_i} \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(\mathbf{s}_{-N}^{\ell-1}|s_\ell, \mathbf{y}_{-N+1}^{\ell})\right), \tag{B.52}$$

and equality $(*)$ follows from Lemma 42. But this is equal to

$$J = \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\ell\in\mathcal{I}_N} \sum_{\substack{\mathbf{s} \\ s_{\ell-1}=i, s_\ell=j}} Q(\mathbf{s}_{-N}^{\ell-2}, \mathbf{s}_{\ell+1}^{N}|s_{\ell-1}, s_\ell) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_{\ell-1}, s_\ell|\mathbf{y})\right) \tag{B.53}$$

$$+ \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\ell\in\mathcal{I}_N} \sum_{\substack{\mathbf{s}_\ell^N \\ s_\ell=j}} Q(\mathbf{s}_{\ell+1}^{N}|s_\ell) \sum_{\mathbf{y}_{\ell+1}^{N}} W(\mathbf{y}_{\ell+1}^{N}|\mathbf{s}_\ell^N) \log\left(V(\mathbf{s}_{\ell+1}^{N}|s_\ell, \mathbf{y}_{\ell+1}^{N})\right) \tag{B.54}$$

$$+ \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\ell\in\mathcal{I}_N} \sum_{\substack{\mathbf{s}_{-N}^{\ell-1} \\ s_{\ell-1}=i}} Q(\mathbf{s}_{-N}^{\ell-2}|s_{\ell-1}) \sum_{\mathbf{y}_{-N+1}^{\ell-1}} W(\mathbf{y}_{-N+1}^{\ell-1}|\mathbf{s}_{-N}^{\ell-1}) \log\left(V(\mathbf{s}_{-N}^{\ell-2}|s_{\ell-1}, \mathbf{y}_{-N+1}^{\ell-1})\right) \tag{B.55}$$

$$- \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\ell\in\mathcal{I}'_N} \sum_{\substack{\mathbf{s} \\ s_\ell=i}} Q(\mathbf{s}_{-N}^{\ell-1}, \mathbf{s}_{\ell+1}^{N}|s_\ell) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_\ell|\mathbf{y})\right) \tag{B.56}$$

$$- \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\ell\in\mathcal{I}'_N} \sum_{\substack{\mathbf{s}_\ell^N \\ s_\ell=i}} Q(\mathbf{s}_{\ell+1}^{N}|s_\ell) \sum_{\mathbf{y}_{\ell+1}^{N}} W(\mathbf{y}_{\ell+1}^{N}|\mathbf{s}_\ell^N) \log\left(V(\mathbf{s}_{\ell+1}^{N}|s_\ell, \mathbf{y}_{\ell+1}^{N})\right) \tag{B.57}$$

$$- \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \sum_{\ell\in\mathcal{I}'_N} \sum_{\substack{\mathbf{s}_{-N}^{\ell} \\ s_\ell=i}} Q(\mathbf{s}_{-N}^{\ell-1}|s_\ell) \sum_{\mathbf{y}_{-N+1}^{\ell}} W(\mathbf{y}_{-N+1}^{\ell}|\mathbf{s}_{-N}^{\ell}) \log\left(V(\mathbf{s}_{-N}^{\ell-1}|s_\ell, \mathbf{y}_{-N+1}^{\ell})\right). \tag{B.58}$$

In the following, we use $\overline{\overline{T}}_{ij}^{(N)}$, $\overline{T}_{i}^{(N)}$, $\overline{\overline{T}}_{ij}^{(N)}(\ell)$, and $\overline{T}_{i}^{(N)}(\ell)$ from (**??**), (3.37), (3.38) and (3.38), which we repeat here for the convience of the reader,[2]

$$T_{ij}^{(N)} \triangleq \overline{\overline{T}}_{ij}^{(N)} - \overline{T}_i^{(N)}, \tag{B.59}$$

$$\begin{cases} \overline{\overline{T}}_{ij}^{(N)} & \triangleq \frac{1}{N'} \sum_{\ell\in\mathcal{I}_N} \overline{\overline{T}}_{ij}^{(N)}(\ell) \\ \overline{T}_i^{(N)} & \triangleq \frac{1}{N'} \sum_{\ell\in\mathcal{I}'_N} \overline{T}_i^{(N)}(\ell) \end{cases} \tag{B.60}$$

$$\begin{cases} \overline{\overline{T}}_{ij}^{(N)}(\ell) & \triangleq \sum_{\substack{\mathbf{s} \\ s_{\ell-1}=i, s_\ell=j}} Q(\mathbf{s}_{-N}^{\ell-2}, \mathbf{s}_{\ell+1}^{N} | s_{\ell-1}, s_\ell) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_{\ell-1}, s_\ell|\mathbf{y})\right) \quad (\text{for } \ell\in\mathcal{I}_N) \\ \overline{T}_i^{(N)}(\ell) & \triangleq \sum_{\substack{\mathbf{s} \\ s_\ell=i}} Q(\mathbf{s}_{-N}^{\ell-1}, \mathbf{s}_{\ell+1}^{N} | s_\ell) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_\ell|\mathbf{y})\right) \quad (\text{for } \ell\in\mathcal{I}'_N) \end{cases} \tag{B.61}$$

and we define additionally

$$\overrightarrow{\chi}_i(\ell) \triangleq \frac{1}{N'} \sum_{\substack{\mathbf{s}_\ell^N \\ s_\ell=i}} Q(\mathbf{s}_{\ell+1}^N|s_\ell) \sum_{\mathbf{y}_{\ell+1}^N} W(\mathbf{y}_{\ell+1}^N|\mathbf{s}_\ell^N) \log\left(V(\mathbf{s}_{\ell+1}^N|s_\ell, \mathbf{y}_{\ell+1}^N)\right), \tag{B.62}$$

$$\overleftarrow{\chi}_i(\ell) \triangleq \frac{1}{N'} \sum_{\substack{\mathbf{s}_{-N}^\ell \\ s_\ell=i}} Q(\mathbf{s}_{-N}^{\ell-1}|s_\ell) \sum_{\mathbf{y}_{-N+1}^\ell} W(\mathbf{y}_{-N+1}^\ell|\mathbf{s}_{-N}^\ell) \log\left(V(\mathbf{s}_{-N}^{\ell-1}|s_\ell, \mathbf{y}_{-N+1}^\ell)\right), \tag{B.63}$$

$$\overrightarrow{\chi}_i \triangleq \sum_{\ell\in\mathcal{I}'_N} \overrightarrow{\chi}_i(\ell) \quad (\text{for all } i). \tag{B.64}$$

These expressions help us to analyze the above sum representing J.

- The first term is equal to $(1/N')\sum_{(i,j)\in\mathcal{A}} Q_{ij}^\alpha \sum_{\ell\in\mathcal{I}_N} \overline{\overline{T}}_{ij}^{(N)}(\ell) = \sum_{(i,j)\in\mathcal{A}} Q_{ij}^\alpha \overline{\overline{T}}_{ij}^{(N)}$,
- the second term is equal to $\sum_{(i,j)\in\mathcal{A}} Q_{ij}^\alpha \sum_{\ell\in\mathcal{I}_N} \overrightarrow{\chi}_j(\ell)$,
- the third term is equal to $\sum_{(i,j)\in\mathcal{A}} Q_{ij}^\alpha \sum_{\ell\in\mathcal{I}_N} \overleftarrow{\chi}_i(\ell-1)$,
- the fourth term is equal to $-(1/N')\sum_{(i,j)\in\mathcal{A}} Q_{ij}^\alpha \sum_{\ell\in\mathcal{I}'_N} \overline{T}_i(\ell) = -\sum_{(i,j)\in\mathcal{A}} Q_{ij}^\alpha \overline{T}_i^{(N)}$,
- the fifth term is equal to $-\sum_{(i,j)\in\mathcal{A}} Q_{ij}^\alpha \sum_{\ell\in\mathcal{I}'_N} \overrightarrow{\chi}_i(\ell)$,
- the sixth term is equal to $-\sum_{(i,j)\in\mathcal{A}} Q_{ij}^\alpha \sum_{\ell\in\mathcal{I}'_N} \overleftarrow{\chi}_i(\ell)$.

[2]Note that in $T_{ij}^{(N)} = \overline{\overline{T}}_{ij}^{(N)} - \overline{T}_i^{(N)}$, every state s_ℓ for $\ell\in[-N,N]$ is "estimated" (i.e. appearing in $V(.|\mathbf{y})$ before the bar) once (when counting with "multiplicities").

Therefore, we can rewrite J as

$$J = \sum_{(i,j)\in\mathcal{A}} Q^{\alpha}_{ij} \left[\overline{\overline{T}}^{(N)}_{ij} - \overline{T}^{(N)}_{i} + \sum_{\ell\in\mathcal{I}_N} \overrightarrow{\chi}_j(\ell) + \sum_{\ell\in\mathcal{I}_N} \overleftarrow{\chi}_i(\ell-1) - \sum_{\ell\in\mathcal{I}'_N} \overrightarrow{\chi}_i(\ell) - \sum_{\ell\in\mathcal{I}'_N} \overleftarrow{\chi}_i(\ell) \right] \tag{B.65}$$

$$= \sum_{(i,j)\in\mathcal{A}} Q^{\alpha}_{ij} \left[T^{(N)}_{ij} + \sum_{\ell=-N+1}^{N} \overrightarrow{\chi}_j(\ell) + \sum_{\ell=-N}^{N-1} \overleftarrow{\chi}_i(\ell) - \sum_{\ell=-N+1}^{N-1} \overrightarrow{\chi}_i(\ell) - \sum_{\ell=-N+1}^{N-1} \overleftarrow{\chi}_i(\ell) \right] \tag{B.66}$$

$$\stackrel{(*)}{=} \sum_{(i,j)\in\mathcal{A}} Q^{\alpha}_{ij} \left[T^{(N)}_{ij} + \sum_{\ell=-N+1}^{N-1} \overrightarrow{\chi}_j(\ell) - \sum_{\ell=-N+1}^{N-1} \overrightarrow{\chi}_i(\ell) \right] \tag{B.67}$$

$$\stackrel{(**)}{=} \sum_{(i,j)\in\mathcal{A}} Q^{\alpha}_{ij} \left[T^{(N)}_{ij} + \overrightarrow{\chi}_j - \overrightarrow{\chi}_i \right], \tag{B.68}$$

where equality $(*)$ follows from $\overrightarrow{\chi}_i(N) = 0$ and $\overleftarrow{\chi}_i(-N) = 0$, and equality $(**)$ follows from (B.64). The side result

$$\sum_{(i,j)\in\mathcal{A}} (\overrightarrow{\chi}_j - \overrightarrow{\chi}_i) \cdot Q^{\alpha}_{ij} = \left(\sum_{(i,j)\in\mathcal{A}} \overrightarrow{\chi}_j \cdot Q^{\alpha}_{ij} \right) - \left(\sum_{i\in\mathcal{S}} \overrightarrow{\chi}_i \cdot \sum_{j\in\overrightarrow{\mathcal{A}_i}} Q^{\alpha}_{ij} \right) \tag{B.69}$$

$$\stackrel{(*)}{=} \left(\sum_{(i,j)\in\mathcal{A}} \overrightarrow{\chi}_j \cdot Q^{\alpha}_{ij} \right) - \left(\sum_{i\in\mathcal{S}} \overrightarrow{\chi}_i \cdot \sum_{k\in\overleftarrow{\mathcal{A}_i}} Q^{\alpha}_{ki} \right) \tag{B.70}$$

$$= \left(\sum_{(i,j)\in\mathcal{A}} \overrightarrow{\chi}_j \cdot Q^{\alpha}_{ij} \right) - \left(\sum_{(k,i)\in\mathcal{A}} \overrightarrow{\chi}_i Q^{\alpha}_{ki} \right) \tag{B.71}$$

$$\stackrel{(**)}{=} \left(\sum_{(i,j)\in\mathcal{A}} \overrightarrow{\chi}_j \cdot Q^{\alpha}_{ij} \right) - \left(\sum_{(i',j')\in\mathcal{A}} \overrightarrow{\chi}_{j'} \cdot Q^{\alpha}_{i'j'} \right) = 0, \tag{B.72}$$

where equality in $(*)$ follows from $\sum_{j\in\overrightarrow{\mathcal{A}_i}} Q^{\alpha}_{ij} = \sum_{k\in\overleftarrow{\mathcal{A}_i}} Q^{\alpha}_{ki}$ (for all $i \in \mathcal{S}$), and at equality in $(**)$ we have made the substitutions $j' \triangleq i$ and $i' \triangleq k$, helps us to simplify (B.68) even more to

$$J = \sum_{(i,j)\in\mathcal{A}} Q^{\alpha}_{ij} \cdot T^{(N)}_{ij}. \tag{B.73}$$

Remembering that we set $J \triangleq -\frac{\mathrm{d}}{\mathrm{d}\alpha} f^{(N)}_4(\alpha, W)$, we have the desired result.

B.6 Proof of Lemma 28

This follows easily from Def. 27 and Lemma 24.

B.7 Proof of Lemma 29

From Def. 27 we have for some fixed $\tilde{\alpha}$

$$f_4'^{(N)}(\tilde{\alpha}, \alpha, W) = -\sum_{(i,j)\in\mathcal{A}} Q_{ij}(\alpha) \cdot T_{ij}^{(N)}(\tilde{\alpha}). \tag{B.74}$$

Because the $T_{ij}^{(N)}(\tilde{\alpha})$'s are independent of α, we easily get

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_4'^{(N)}(\tilde{\alpha}, \tilde{\alpha}, W)\bigg|_{\alpha=\tilde{\alpha}} = -\frac{\mathrm{d}}{\mathrm{d}\alpha} \sum_{(i,j)\in\mathcal{A}} Q_{ij}(\alpha) \cdot T_{ij}^{(N)}(\tilde{\alpha})\Bigg|_{\alpha=\tilde{\alpha}} = -\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha}(\alpha) \cdot T_{ij}^{(N)}(\tilde{\alpha})\Bigg|_{\alpha=\tilde{\alpha}} \tag{B.75}$$

$$= -\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha}(\tilde{\alpha}) \cdot T_{ij}^{(N)}(\tilde{\alpha}). \tag{B.76}$$

This expression is equivalent to $\frac{\mathrm{d}}{\mathrm{d}\alpha} f_4^{(N)}(\alpha, W)$ (as given in Lemma 25) evaluated at $\alpha = \tilde{\alpha}$.

B.8 Proof of Theorem 31

From Def. 19 we have

$$I^{(N)}(Q_{ij}, W) = f_1^{(N)}(Q_{ij}) - f_4^{(N)}(Q_{ij}, W). \tag{B.77}$$

Using Lemmas 22 and 24 this turns into

$$I^{(N)}(Q_{ij}, W) = -\sum_{(i,j)\in\mathcal{A}} Q_{ij} \log(p_{ij}) - \frac{1}{N'} \sum_{i\in\mathcal{S}} \mu_i \log(\mu_i) + \sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot T_{ij}^{(N)} \tag{B.78}$$

$$\stackrel{(*)}{=} \sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \left[-\log(p_{ij}) - \frac{1}{N'} \mu_i + T_{ij}^{(N)} \right], \tag{B.79}$$

where at equality $(*)$ we used that $\sum_{i\in\mathcal{S}} \mu_i = \sum_{(i,j)\in\mathcal{A}} Q_{ij}$.

B.9 Proof of Theorem 32

From Def. 19 we have

$$I^{(N)}(Q_{ij}, W) = f_1^{(N)}(Q_{ij}) - f_4^{(N)}(Q_{ij}, W), \tag{B.80}$$

therefore

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} I^{(N)}(Q_{ij}, W) = \frac{\mathrm{d}}{\mathrm{d}\alpha} f_1^{(N)}(Q_{ij}) - \frac{\mathrm{d}}{\mathrm{d}\alpha} f_4^{(N)}(Q_{ij}, W) \tag{B.81}$$

$$\stackrel{(*)}{=} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \cdot \left[-\log(p_{ij}) - \frac{1}{N'} \log(\mu_i) \right] + \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha}(\alpha) \cdot T_{ij}^{(N)} \tag{B.82}$$

$$= \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \cdot \left[-\log(p_{ij}) - \frac{1}{N'} \log(\mu_i) + T_{ij}^{(N)} \right], \tag{B.83}$$

where equality $(*)$ follows from Lemmas 23 and 25.

B.10 Proof of Definition 33

We calculate

$$\Psi^{(N)}(\tilde{Q}_{ij}, Q_{ij}, W) \stackrel{\triangle}{=} f_1^{(N)}(Q_{ij}) - f_4'^{(N)}(\tilde{Q}_{ij}, Q_{ij}, W) \tag{B.84}$$

$$\stackrel{(*)}{=} \sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \left[-\log(p_{ij}) - \frac{1}{N'} \log(\mu_i) \right] + \sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \tilde{T}_{ij}^{(N)} \tag{B.85}$$

$$= \sum_{(i,j)\in\mathcal{A}} Q_{ij} \cdot \left[-\log(p_{ij}) - \frac{1}{N'} \log(\mu_i) + \tilde{T}_{ij}^{(N)} \right], \tag{B.86}$$

where equality in $(*)$ follows from Lemma 22 and Def. 27.

B.11 Proof of Theorem 34

In Def. 33 we defined

$$\Psi^{(N)}(\tilde{Q}_{ij}, Q_{ij}, W) \stackrel{\triangle}{=} f_1^{(N)}(Q_{ij}) - f_4'^{(N)}(\tilde{Q}_{ij}, Q_{ij}, W), \tag{B.87}$$

whereas in Def. 19 we had

$$I(Q_{ij}, W) = f_1^{(N)}(Q_{ij}) - f_4^{(N)}(Q_{ij}, W). \tag{B.88}$$

Using the relation $f_4'^{(N)}(\tilde{Q}_{ij}, \tilde{Q}_{ij}, W) = f_4^{(N)}(\tilde{Q}_{ij}, W)$, which was shown to hold in Lemma 28, we get the desired result.

B.12 Proof of Theorem 35

Let $Q_{ij} = Q_{ij}(\alpha)$, fix some $\tilde{\alpha}$, and let $\tilde{T}_{ij}^{(N)} = T_{ij}^{(N)}(Q_{ij}(\alpha), W)$. We have to derive

$$\Psi^{(N)}\big(\tilde{\alpha}, \alpha, W\big) \stackrel{\triangle}{=} \sum_{(i,j)\in\mathcal{A}} Q_{ij}(\alpha) \left[-\log\left(p_{ij}(\alpha)\right) + \tilde{T}_{ij}^{(N)} \right], \tag{B.89}$$

with respect to α. This is not too difficult, because here the $\tilde{T}_{ij}^{(N)}$'s are constants.

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \Psi^{(N)}\big(Q_{ij}(\tilde{\alpha}), Q_{ij}(\alpha), W\big) \tag{B.90}$$

$$= \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha}(\alpha) \left[-\log\left(p_{ij}(\alpha)\right) + \tilde{T}_{ij}^{(N)} \right] + \sum_{(i,j)\in\mathcal{A}} Q_{ij}(\alpha) \frac{1}{p_{ij}(\alpha)} p_{ij}^{\alpha}(\alpha) \tag{B.91}$$

$$= \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha}(\alpha) \left[-\log\left(p_{ij}(\alpha)\right) + \tilde{T}_{ij}^{(N)} \right] + \sum_{i\in\mathcal{S}} \mu_i(\alpha) \frac{\mathrm{d}}{\mathrm{d}\alpha} \sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij}(\alpha) \tag{B.92}$$

$$\stackrel{(*)}{=} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha}(\alpha) \left[-\log\left(p_{ij}(\alpha)\right) + \tilde{T}_{ij}^{(N)} \right], \tag{B.93}$$

where equality $(*)$ follows from $\sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij}(\alpha) = 1$ for any α. Evaluating (B.93) for $\alpha = \tilde{\alpha}$ and using Th. 31 gives the desired result.

B.13 Proof of Theorem 36

We have to maximize

$$\Psi\left(\{\tilde{Q}_{ij}\},\{Q_{ij}\},W\right)=\sum_{(i,j)\in\mathcal{A}}Q_{ij}\left(\log\left(\frac{\sum_{j'\in\overrightarrow{\mathcal{A}_i}}Q_{ij'}}{Q_{ij}}\right)+\tilde{T}_{ij}\right) \tag{B.94}$$

over $\{Q_{ij}\}$ under the constraints[3]

$$\sum_{(i,j)\in\mathcal{A}}Q_{ij}=1, \tag{B.95}$$

$$\sum_{k\in\overleftarrow{\mathcal{A}_i}}Q_{ki}=\sum_{j\in\overrightarrow{\mathcal{A}_i}}Q_{ij}\quad\text{(for all } i\in\mathcal{S}\text{)}. \tag{B.96}$$

This is equivalent to setting the gradient of the Lagrangian

$$L=-\sum_{(i,j)\in\mathcal{A}}Q_{ij}\left(\log\left(\frac{\sum_{j'\in\overrightarrow{\mathcal{A}_i}}Q_{ij'}}{Q_{ij}}\right)+\tilde{T}_{ij}\right) \tag{B.97}$$

$$+\lambda\left(\sum_{(i,j)\in\mathcal{A}}Q_{ij}-1\right)+\sum_{i\in\mathcal{S}}\lambda_i\left(\sum_{k\in\overleftarrow{\mathcal{A}_i}}Q_{ki}-\sum_{j\in\overrightarrow{\mathcal{A}_i}}Q_{ij}\right) \tag{B.98}$$

$$=-\sum_{(i,j)\in\mathcal{A}}Q_{ij}\left(\log\left(\frac{\sum_{j'\in\overrightarrow{\mathcal{A}_i}}Q_{ij'}}{Q_{ij}}\right)+\tilde{T}_{ij}\right) \tag{B.99}$$

$$+\lambda\left(\sum_{(i,j)\in\mathcal{A}}Q_{ij}-1\right)+\left(\sum_{(i,j)\in\mathcal{A}}\lambda_jQ_{ij}\right)-\left(\sum_{(i,j)\in\mathcal{A}}\lambda_iQ_{ij}\right) \tag{B.100}$$

equal to zero:

$$\begin{cases}\frac{\partial L}{\partial Q_{ij}} & \overset{!}{=}0\quad\text{(for all } (i,j)\in\mathcal{A}\text{)},\\ \frac{\partial L}{\partial\lambda} & \overset{!}{=}0,\\ \frac{\partial L}{\partial\lambda_i} & \overset{!}{=}0\quad\text{(for all } i\in\mathcal{S}\text{)}.\end{cases} \tag{B.101}$$

Solving these equations we get

$$0\overset{!}{=}\frac{\partial L}{\partial Q_{ij}}=\left(\log\left(\frac{\sum_{j'\in\overrightarrow{\mathcal{A}_i}}Q_{ij'}}{Q_{ij}}\right)+\tilde{T}_{ij}\right)+\underbrace{\sum_{j\in\overrightarrow{\mathcal{A}_i}}Q_{ij}\frac{1}{\sum_{j'\in\overrightarrow{\mathcal{A}_i}}Q_{ij'}}}_{=1}-Q_{ij}\frac{1}{Q_{ij}}+\lambda+\lambda_j-\lambda_i \tag{B.102}$$

$$=-\log p_{ij}+\tilde{T}_{ij}+\lambda+\lambda_j-\lambda_i\quad\text{(for all } (i,j)\in\mathcal{A}\text{)}, \tag{B.103}$$

$$0\overset{!}{=}\frac{\partial L}{\partial\lambda}=\sum_{(i,j)\in\mathcal{A}}Q_{ij}-1, \tag{B.104}$$

$$0\overset{!}{=}\frac{\partial L}{\partial\lambda_i}=\sum_{k\in\overleftarrow{\mathcal{A}_i}}Q_{ki}-\sum_{j\in\overrightarrow{\mathcal{A}_i}}Q_{ij}\quad\text{(for all } i\in\mathcal{S}\text{)}. \tag{B.105}$$

[3]For the moment, we neglect the constraints $Q_{ij}\geq 0$ for all $(i,j)\in\mathcal{A}$.

So

$$p_{ij} = \mathrm{e}^{\lambda_j - \lambda_i + \lambda + \tilde{T}_{ij}} \quad \text{(for all } (i,j) \in \mathcal{A}). \tag{B.106}$$

By the definition of the p_{ij}'s we must have

$$\sum_{j \in \overrightarrow{\mathcal{A}_i}} p_{ij} = 1 \quad \text{(for all } i \in \mathcal{S}), \tag{B.107}$$

therefore we get

$$\sum_{j \in \overrightarrow{\mathcal{A}_i}} \mathrm{e}^{\tilde{T}_{ij}} \mathrm{e}^{\lambda_j} = \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda_i} \quad \text{(for all } i \in \mathcal{S}). \tag{B.108}$$

Let $\tilde{\mathbf{A}}$ be the matrix with entries

$$\tilde{a}_{ij} \triangleq \begin{cases} \mathrm{e}^{\tilde{T}_{ij}} & \text{(if } (i,j) \in \mathcal{A}) \\ 0 & \text{(otherwise)} \end{cases}, \tag{B.109}$$

and let $\mathbf{b}$ be the (row) vector with entries $b_i = \mathrm{e}^{\lambda_i}$, and $\rho = \mathrm{e}^{-\lambda}$, then

$$\tilde{\mathbf{A}} \mathbf{b}^T = \rho \cdot \mathbf{b}^T, \tag{B.110}$$

i.e., $\mathbf{b}^T$ must be a *right* eigenvector of $\tilde{\mathbf{A}}$ with a positive (and therefore also real) eigenvalue ρ. Moreover, all entries of $\mathbf{b}^T$ must be positive (see also the comment on page 61). Inserting these results into (B.106) we get

$$p_{ij} = \frac{b_j}{b_i} \cdot \frac{\tilde{a}_{ij}}{\rho} \quad \text{(for all } (i,j) \in \mathcal{A}), \tag{B.111}$$

From $\sum_{i \in \mathcal{S}} \mu_i p_{ij} = \mu_j$ (for all $j \in \mathcal{S}$) follows

$$\sum_{i \in \mathcal{S}} \mu_i \frac{b_j \tilde{a}_{ij}}{b_i \rho} = \mu_j \quad \text{(for all } j \in \mathcal{S}), \tag{B.112}$$

and by letting the (row) vector $\mathbf{c}$ have entries $c_i = \mu_i/(K b_i)$ (where K will be determined later) we obtain

$$\sum_{i \in \mathcal{S}} c_i \tilde{a}_{ij} = \rho \cdot c_j \quad \text{(for all } j \in \mathcal{S}), \quad \text{or, equivalently,} \quad \mathbf{c}\tilde{\mathbf{A}} = \rho \cdot \mathbf{c}, \tag{B.113}$$

i.e. $\mathbf{c}$ is a *left* eigenvector of $\tilde{\mathbf{A}}$ with eigenvalue ρ, whose entries must be non-negative. Consequently, to fulfill $\sum_{i \in \mathcal{S}} \mu_i = 1$, we must have

$$\mu_i = K \cdot c_i \cdot b_i, \quad \text{(for all } i \in \mathcal{S}), \quad \text{with} \quad K = \frac{1}{\sum_{i \in \mathcal{S}} c_i b_i}. \tag{B.114}$$

Finally, we set $\mu_i^* \triangleq \mu_i$, $p_{ij}^* \triangleq p_{ij}$, and $Q_{ij}^* \triangleq \mu_i^* p_{ij}^*$. We still have to determine what eigenvalue of $\tilde{\mathbf{A}}$ we have to take.

We would now like to show that $\Psi\left(\tilde{Q}_{ij}, Q^*_{ij}, W\right) = \log(\rho)$. This indeedfollows from

$$
\begin{aligned}
&\Psi\left(\{\tilde{Q}_{ij}\}, \{Q^*_{ij}\}, W\right) && \text{(B.115)}\\
&= \sum_{(i,j)\in\mathcal{A}} \mu^*_i p^*_{ij} \left(\log\left(\frac{1}{p^*_{ij}}\right) + \tilde{T}_{ij}\right) && \text{(B.116)}\\
&\overset{(*)}{=} \sum_{(i,j)\in\mathcal{A}} \mu^*_i p_{ij} \left(\log\left(\frac{1}{p^*_{ij}}\right) + \log(\tilde{a}_{ij})\right) && \text{(B.117)}\\
&= \sum_{(i,j)\in\mathcal{A}} kc_i b_i \frac{b_j \tilde{a}_{ij}}{b_i \rho} \left(\log\left(\frac{b_i \rho}{b_j \tilde{a}_{ij}}\right) + \log(\tilde{a}_{ij})\right) && \text{(B.118)}\\
&= \frac{1}{\rho} \sum_{(i,j)\in\mathcal{A}} kc_i \tilde{a}_{ij} b_j \big(\log(b_i) + \log(\rho) - \log(b_j)\big) && \text{(B.119)}\\
&= \frac{1}{\rho}\left(\sum_{i\in\mathcal{S}} kc_i \big(\log(b_i) + \log(\rho)\big) \underbrace{\sum_{j\in\mathcal{S}} \tilde{a}_{ij} b_j}_{=\rho b_i}\right) - \frac{1}{\rho}\left(\sum_{j\in\mathcal{S}} kb_j \log(b_j) \underbrace{\sum_{i\in\overleftarrow{\mathcal{A}}_j} c_i \tilde{a}_{ij}}_{=\rho c_j}\right) && \text{(B.120)}\\
&= \sum_{i\in\mathcal{S}} \mu_i \big(\log(b_i) + \log(\rho)\big) - \sum_{j\in\mathcal{S}} \mu_j \log(b_j) && \text{(B.121)}\\
&= \log(\rho), && \text{(B.122)}
\end{aligned}
$$

where at step $(*)$ we used $\tilde{T}_{ij} = \log(a_{ij})$ for $(i,j) \in \mathcal{A}$.

$\log(\rho)$ would clearly be maximized by taking ρ to be the largest real eigenvalue of $\tilde{\mathbf{A}}$. But, as we have seen before, the right eigenvector corresponding to the eigenvalue ρ must have positive entries and the left eigenvector must have non-negative entries. The question is whether this can be fulfilled at all.

For an irreducible and non-negative matrix $\tilde{\mathbf{A}}$ one can indeed show that these conditions can be met [17], p. 508. One can show that such matrices have a real eigenvalue whose modulus is the largest of all eigenvalues. Moreover, it is an algebraically and geometrically single eigenvalue. (There may be other complex vector having the same modulus, though.) Such an eigenvalue, which is called the *Perron eigenvalue*, has a left and a right eigenvector whose entries are all positive, respectively. When their entries sum to one, respectively, one calls these eigenvectors the *right* and the *left Perron eigenvector*, respectively.

We come now shortly back to the comment in Footnote 3: from the above comments we must have $Q^{(*)}_{ij} \geq 0$ automatically, i.e. neglecting these constraints at the first place was legal.

We now confirm that $\log(\rho)$ is indeed the largest possible value for $\Psi(\{\tilde{Q}_{ij}\}, \{Q_{ij}\}, W)$ for given $\{\tilde{Q}_{ij}\}$ and W and varying Q_{ij}. Let $\{p^*_{ij}\}$ be the solution given in (B.111). For any $\{p_{ij}\}$ with corresponding stationary probabilities $\{\mu_i\}$ we have

$$\Psi\left(\{\tilde{Q}_{ij}\},\{Q^*_{ij}\},W\right) - \Psi\left(\{\tilde{Q}_{ij}\},\{Q_{ij}\},W\right) \tag{B.123}$$

$$= \log(\rho) - \sum_{(i,j)\in\mathcal{A}} \mu_i p_{ij}\left(\log\left(\frac{1}{p_{ij}}\right) + \tilde{T}_{ij}\right) \tag{B.124}$$

$$= \sum_{(i,j)\in\mathcal{A}} \mu_i p_{ij} \log\left(\frac{p_{ij}}{e^{\tilde{T}_{ij}}/\rho}\right) \tag{B.125}$$

$$\stackrel{(*)}{=} \sum_{i\in\mathcal{S}} \mu_i \underbrace{\sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij}\log\left(\frac{p_{ij}}{p^*_{ij}}\right)}_{\geq 0\,(**)} + \sum_{(i,j)\in\mathcal{A}} \mu_i p_{ij}\log(b_j) - \sum_{(i,j)\in\mathcal{A}} \mu_i p_{ij}\log(b_i) \tag{B.126}$$

$$\geq \sum_{j\in\mathcal{S}} \log(b_j) \underbrace{\sum_{i\in\overleftarrow{\mathcal{A}_j}} \mu_i p_{ij}}_{=\mu_j} - \sum_{i\in\mathcal{S}} \mu_i \log(b_i) \underbrace{\sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij}}_{=1} \tag{B.127}$$

$$= \sum_{j\in\mathcal{S}} \mu_j \log(b_j) - \sum_{i\in\mathcal{S}} \mu_i \log(b_i) = 0, \tag{B.128}$$

where at step $(*)$ we used the fact that for $a_{ij} = 1$ we have $e^{\tilde{T}_{ij}}/\rho = p^*_{ij}\cdot b_i/b_j$ and $(**)$ follows from the fact that relative entropies are non-negative.

We note that once given the correct solution, (B.123) - (B.128) are sufficient to show that this is also the optimal solution.

B.14 Proof of Theorem 38

Assume that at itertion r we found $\{Q^{\langle r\rangle}_{ij}\}$ and that this is a stationary point of Alg. 37. For any parametrization $Q_{ij} = Q_{ij}(\alpha)$ with a single parameter α where $Q^{\langle r\rangle}_{ij} = Q_{ij}(\tilde{\alpha})$ (for all $(i,j)\in\mathcal{A}$) for some $\tilde{\alpha}$, we must have

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\Psi\left(\tilde{\alpha},\alpha,W\right)\bigg|_{\alpha=\tilde{\alpha}} = 0. \tag{B.129}$$

But by Th. 35 we have

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\Psi\left(\tilde{\alpha},\alpha,W\right)\bigg|_{\alpha=\tilde{\alpha}} = \frac{\mathrm{d}}{\mathrm{d}\alpha}I(\alpha,W)\bigg|_{\alpha=\tilde{\alpha}}, \tag{B.130}$$

therefore

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}I(\alpha,W)\bigg|_{\alpha=\tilde{\alpha}} = 0. \tag{B.131}$$

We can also show the reverse dirction, from which we conclude that the stationary points of Alg. 37 correspond one-to-one to critical points of the information rate curve.

But critical points of the information rate curve that are not maxima are not stable stationary points of Alg. 37. This is because for estimating the T_{ij}'s we take only finite-length state and output sequences, therefore at each iteration the estimates vary slightly. At a local minima, as soon as the algorithm gives a $\{Q_{ij}\}$ that does not correspond to the critical point, the algorithm will with probability 1 (as $N \to \infty$) not return to the minimum point at subsequent iterations. At terrace points the algorithm will with non-vanishing probability give a new $\{Q_{ij}\}$ whose information rate is larger; in subsequent iterations the algorithm will with probability 1 (as $N \to \infty$) not return to the terrace point.

B.15 Proof of Remark 39

We consider the first proposed possibility to compute T_{ij}. From Def. 21,

$$\overline{\overline{T}}_{ij}^{(N)} = \frac{1}{N'} \sum_{\ell \in \mathcal{I}_N} \sum_{\substack{\mathbf{s} \\ s_{\ell-1}=i, s_\ell=j}} Q(\mathbf{s}_{-N}^{\ell-2}, \mathbf{s}_{\ell+1}^{N} | s_{\ell-1}, s_\ell) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_{\ell-1}, s_\ell|\mathbf{y})\right) \tag{B.132}$$

$$= \frac{1}{N'} \sum_{\ell \in \mathcal{I}_N} \sum_{\substack{\mathbf{s} \\ s_{\ell-1}=i, s_\ell=j}} \frac{Q(\mathbf{s})}{Q_{ij}} \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_{\ell-1}, s_\ell|\mathbf{y})\right) \tag{B.133}$$

$$= \sum_{\mathbf{s}} \sum_{\mathbf{y}} Q(\mathbf{s}) W(\mathbf{y}|\mathbf{s}) \left[\frac{1}{N' Q_{ij}} \sum_{\substack{\ell \in \mathcal{I}_N \\ s_{\ell-1}=i, s_\ell=j}} \log\left(V(s_{\ell-1}, s_\ell|\mathbf{y})\right) \right] \tag{B.134}$$

Let $\check{\mathbf{s}}$ be a (typical) state sequence and $\check{\mathbf{y}}$ be a (typical) output sequence. Then we have the approximation

$$\check{\overline{\overline{T}}}_{ij}^{(N)} \approx \frac{1}{N' Q_{ij}} \sum_{\substack{\ell \in \mathcal{I}_N \\ \check{s}_{\ell-1}=i, \check{s}_\ell=j}} \log\left(V_{\ell-1,\ell}(i, j|\check{\mathbf{y}})\right), \tag{B.135}$$

$$\tag{B.136}$$

for finite N, and we have equality with probability 1 for $N \to \infty$. Similarly,

$$\overline{\overline{T}}_{ij}^{(N)} = \frac{1}{N'} \sum_{\ell \in \mathcal{I}'_N} \sum_{\substack{\mathbf{s} \\ s_\ell=i}} Q(\mathbf{s}_{-N}^{\ell-1}, \mathbf{s}_{\ell+1}^{N} | s_\ell) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_\ell|\mathbf{y})\right) \tag{B.137}$$

$$= \frac{1}{N'} \sum_{\ell \in \mathcal{I}'_N} \sum_{\substack{\mathbf{s} \\ s_\ell=i}} \frac{Q(\mathbf{s})}{\mu_i} \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log\left(V(s_\ell|\mathbf{y})\right) \tag{B.138}$$

$$= \sum_{\mathbf{s}} \sum_{\mathbf{y}} Q(\mathbf{s}) W(\mathbf{y}|\mathbf{s}) \left[\frac{1}{N' \mu_i} \sum_{\substack{\ell \in \mathcal{I}'_N \\ s_\ell=i}} \log\left(V(s_\ell|\mathbf{y})\right) \right] \tag{B.139}$$

Let $\check{\mathbf{s}}$ be a (typical) state sequence and $\check{\mathbf{y}}$ be a (typical) output sequence. Then we have the approximation

$$\check{\overline{T}}_{i}^{(N)} \approx \frac{1}{N' \mu_i} \sum_{\substack{\ell \in \mathcal{I}'_N \\ \check{s}_\ell=i}} \log\left(V_\ell(i|\check{\mathbf{y}})\right) \tag{B.140}$$

for finite N, and we have equality with probability 1 for $N \to \infty$. Similarly,

We now consider the second proposed possibility to compute T_{ij}. We transform the expression of $\overline{\overline{T}}_{ij}^{(N)}$.

$$\overline{\overline{T}}_{ij}^{(N)} = \frac{1}{N'} \sum_{\ell \in \mathcal{I}_N} \sum_{\substack{\mathbf{s} \\ s_{\ell-1}=i, s_\ell=j}} Q(\mathbf{s}_{-N}^{\ell-2}, \mathbf{s}_{\ell+1}^{N} | s_{\ell-1}, s_\ell) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log \left(V(s_{\ell-1}, s_\ell | \mathbf{y}) \right) \tag{B.141}$$

$$= \sum_{\mathbf{y}} R(\mathbf{y}) \frac{1}{N'} \sum_{\ell \in \mathcal{I}_N} \sum_{\substack{\mathbf{s} \\ s_{\ell-1}=i, s_\ell=j}} \frac{Q(\mathbf{s}_{-N}^{\ell-2}, \mathbf{s}_{\ell+1}^{N} | s_{\ell-1}, s_\ell) W(\mathbf{y}|\mathbf{s})}{R(\mathbf{y})} \log \left(V(s_{\ell-1}, s_\ell | \mathbf{y}) \right) \tag{B.142}$$

$$= \sum_{\mathbf{y}} R(\mathbf{y}) \frac{1}{N'} \sum_{\ell \in \mathcal{I}_N} \sum_{\substack{\mathbf{s} \\ s_{\ell-1}=i, s_\ell=j}} \frac{Q(\mathbf{s}) W(\mathbf{y}|\mathbf{s})}{Q_{ij} R(\mathbf{y})} \log \left(V(s_{\ell-1}, s_\ell | \mathbf{y}) \right) \tag{B.143}$$

$$= \sum_{\mathbf{y}} R(\mathbf{y}) \frac{1}{N'} \sum_{\ell \in \mathcal{I}_N} \sum_{\substack{\mathbf{s} \\ s_{\ell-1}=i, s_\ell=j}} \frac{V(\mathbf{s}|\mathbf{y})}{Q_{ij}} \log \left(V(s_{\ell-1}, s_\ell | \mathbf{y}) \right) \tag{B.144}$$

$$= \sum_{\mathbf{y}} R(\mathbf{y}) \left[\frac{1}{N'} \sum_{\ell \in \mathcal{I}_N} \frac{V_{\ell-1,\ell}(i,j|\mathbf{y})}{Q_{ij}} \log \left(V_{\ell-1,\ell}(i,j|\mathbf{y}) \right) \right], \tag{B.145}$$

Let $\check{\mathbf{y}}$ be a (typical) output sequence. Then we have the approximation

$$\check{\overline{\overline{T}}}_{ij}^{(N)} \approx \frac{1}{N'} \sum_{\ell \in \mathcal{I}_N} \frac{V_{\ell-1,\ell}(i,j|\check{\mathbf{y}})}{Q_{ij}} \log \left(V_{\ell-1,\ell}(i,j|\check{\mathbf{y}}) \right) \tag{B.146}$$

for finite N, and we have equality with probability 1 for $N \to \infty$. Similarly,

$$\overline{T}_{i}^{(N)} = \frac{1}{N'} \sum_{\ell \in \mathcal{I}'_N} \sum_{\substack{\mathbf{s} \\ s_\ell=i}} Q(\mathbf{s}_{-N}^{\ell-1}, \mathbf{s}_{\ell+1}^{N} | s_\ell) \sum_{\mathbf{y}} W(\mathbf{y}|\mathbf{s}) \log \left(V(s_\ell | \mathbf{y}) \right) \tag{B.147}$$

$$= \sum_{\mathbf{y}} R(\mathbf{y}) \frac{1}{N'} \sum_{\ell \in \mathcal{I}'_N} \sum_{\substack{\mathbf{s} \\ s_\ell=i}} \frac{Q(\mathbf{s}_{-N}^{\ell-1}, \mathbf{s}_{\ell+1}^{N} | s_\ell) W(\mathbf{y}|\mathbf{s})}{R(\mathbf{y})} \log \left(V(s_\ell | \mathbf{y}) \right) \tag{B.148}$$

$$= \sum_{\mathbf{y}} R(\mathbf{y}) \frac{1}{N'} \sum_{\ell \in \mathcal{I}'_N} \sum_{\substack{\mathbf{s} \\ s_\ell=i}} \frac{Q(\mathbf{s}) W(\mathbf{y}|\mathbf{s})}{\mu_i R(\mathbf{y})} \log \left(V(s_\ell | \mathbf{y}) \right) \tag{B.149}$$

$$= \sum_{\mathbf{y}} R(\mathbf{y}) \frac{1}{N'} \sum_{\ell \in \mathcal{I}'_N} \sum_{\substack{\mathbf{s} \\ s_\ell=i}} \frac{V(\mathbf{s}|\mathbf{y})}{\mu_i} \log \left(V(s_\ell | \mathbf{y}) \right) \tag{B.150}$$

$$= \sum_{\mathbf{y}} R(\mathbf{y}) \left[\frac{1}{N'} \sum_{\ell \in \mathcal{I}'_N} \frac{V_\ell(i|\mathbf{y})}{\mu_i} \log \left(V_\ell(i|\mathbf{y}) \right) \right]. \tag{B.151}$$

Let $\check{\mathbf{y}}$ be an (typical) output sequence. Then we have the approximation

$$\check{\overline{T}}_{ij}^{(N)} \approx \frac{1}{N'} \sum_{\ell \in \mathcal{I}'_N} \frac{V_\ell(i|\check{\mathbf{y}})}{Q_{ij}} \log \left(V_\ell(i|\check{\mathbf{y}}) \right) \tag{B.152}$$

for finite N, and we have equality with probability 1 for $N \to \infty$.

Remark: We are aware of the fact, that this section needs more comments concerning the required ergodicity to otain these results.

B.16 Proof of Lemma 40

To prove concavity, we express $Q_{ij} \triangleq Q_{ij}(\alpha)$ as a linear function in a single parameter α. If a function is concave in α for any such parametrization, then the function is concave in $\{Q_{ij}\}$. Therefore we assume

$$Q_{ij}^{\alpha} \triangleq \frac{\mathrm{d}}{\mathrm{d}\alpha} Q_{ij} \quad \text{are constants,} \tag{B.153}$$

$$Q_{ij}^{\alpha\alpha} \triangleq \frac{\mathrm{d}^2}{\mathrm{d}^2\alpha} Q_{ij} = 0. \tag{B.154}$$

Again, we have

$$0 = \frac{\mathrm{d}}{\mathrm{d}\alpha} \sum_{(i,j)\in\mathcal{A}} Q_{ij} = \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha}. \tag{B.155}$$

Note that

$$0 = \frac{\mathrm{d}}{\mathrm{d}\alpha} \sum_{\mathbf{s}} Q(\mathbf{s}) = \sum_{\mathbf{s}} \frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}). \tag{B.156}$$

From Lemma 23 we already have the derivative of $f_1^{(N)}(\alpha)$ with respect to α.

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f_1^{(N)}(\alpha) = -\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \log(p_{ij}) - \frac{1}{N'} \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \log(\mu_i) \tag{B.157}$$

$$\triangleq J_{11} + J_{12}, \tag{B.158}$$

where we defined the first term (with sign) to be J_{11} and the second term (with sign) to be J_{12}. Continuing, we get

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} J_{11} = -\sum_{(i,j)\in\mathcal{A}} \left(Q_{ij}^{\alpha\alpha}\right) \log\left(\frac{Q_{ij}}{\sum_{j'\in\overrightarrow{\mathcal{A}_i}} Q_{ij'}}\right) \tag{B.159}$$

$$-\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \frac{1}{Q_{ij}} Q_{ij}^{\alpha} + \sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha} \frac{1}{\sum_{j'\in\overrightarrow{\mathcal{A}_i}} Q_{ij'}} \sum_{j''} Q_{ij''}^{\alpha} \tag{B.160}$$

$$= -\sum_{(i,j)\in\mathcal{A}} \frac{(Q_{ij}^{\alpha})^2}{Q_{ij}} + \sum_{i\in\mathcal{S}} \frac{\left(\sum_{j\in\overrightarrow{\mathcal{A}_i}} Q_{ij}^{\alpha}\right)^2}{\sum_{j'\in\overrightarrow{\mathcal{A}_i}} Q_{ij'}} \tag{B.161}$$

To proceed, we need the Cauchy-Schwarz inequality which says that

$$\left(\sum_j a_j b_j\right)^2 \le \left(\sum_j a_j^2\right) \cdot \left(\sum_j b_j^2\right), \tag{B.162}$$

where equality holds if and only if $a_j = b_j$ for all j. With

$$a_j = \frac{Q_{ij}^{\alpha}}{\sqrt{Q_{ij}}}, \tag{B.163}$$

$$b_j = \sqrt{Q_{ij}}, \tag{B.164}$$

we get

$$\left(\sum_{j\in\overrightarrow{\mathcal{A}_i}} Q_{ij}^{\alpha}\right)^2 \leq \left(\sum_{j\in\overrightarrow{\mathcal{A}_i}} \frac{(Q_{ij}^{\alpha})^2}{Q_{ij}}\right)\left(\sum_{j\in\overrightarrow{\mathcal{A}_i}} Q_{ij}\right). \tag{B.165}$$

Using this side result we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}J_{11} = -\sum_{(i,j)\in\mathcal{A}} \frac{(Q_{ij}^{\alpha})^2}{Q_{ij}} + \sum_{i\in\mathcal{S}} \frac{\left(\sum_{j\in\overrightarrow{\mathcal{A}_i}} Q_{ij}^{\alpha}\right)^2}{\sum_{j'\in\overrightarrow{\mathcal{A}_i}} Q_{ij'}} \tag{B.166}$$

$$\leq -\sum_{(i,j)\in\mathcal{A}} \frac{(Q_{ij}^{\alpha})^2}{Q_{ij}} + \sum_{i\in\mathcal{S}} \frac{\left(\sum_{j\in\overrightarrow{\mathcal{A}_i}} \frac{(Q_{ij}^{\alpha})^2}{Q_{ij}}\right)\left(\sum_{j\in\overrightarrow{\mathcal{A}_i}} Q_{ij}\right)}{\sum_{j'\in\overrightarrow{\mathcal{A}_i}} Q_{ij'}} \tag{B.167}$$

$$= -\sum_{(i,j)\in\mathcal{A}} \frac{(Q_{ij}^{\alpha})^2}{Q_{ij}} + \sum_{(i,j)\in\mathcal{A}} \frac{(Q_{ij}^{\alpha})^2}{Q_{ij}} = 0. \tag{B.168}$$

On the other hand,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}J_{12} = -\frac{1}{N'}\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha\alpha}\log(\mu_i) - \frac{1}{N'}\sum_{(i,j)\in\mathcal{A}} Q_{ij}^{\alpha}\frac{1}{\mu_i}\mu_i^{\alpha} \tag{B.169}$$

$$\overset{(*)}{=} -\frac{1}{N'}\sum_{(i,j)\in\mathcal{A}} \mu_i^{\alpha}p_{ij}\frac{1}{\mu_i}\mu_i^{\alpha} - \frac{1}{N'}\sum_{(i,j)\in\mathcal{A}} \mu_i p_{ij}^{\alpha}\frac{1}{\mu_i}\mu_i^{\alpha} \tag{B.170}$$

$$= -\frac{1}{N'}\sum_{i\in\mathcal{S}} \frac{(\mu_i^{\alpha})^2}{\mu_i}\sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij} - \frac{1}{N'}\sum_{i\in\mathcal{S}} \mu_i^{\alpha}\sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij}^{\alpha} \tag{B.171}$$

$$\overset{(**)}{=} -\frac{1}{N'}\sum_{i\in\mathcal{S}} \frac{(\mu_i^{\alpha})^2}{\mu_i} \leq 0, \tag{B.172}$$

where equality $(*)$ follows from $Q_{ij}^{\alpha\alpha} = 0$ and $Q_{ij}^{\alpha} = \mu_i^{\alpha}p_{ij} + \mu_i p_{ij}^{\alpha}$ and equality $(**)$ from $\sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij} = 1$ for all $i \in \mathcal{S}$ and $\sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij}^{\alpha} = \frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_{j\in\overrightarrow{\mathcal{A}_i}} p_{ij} = 0$. Combining, we get

$$\frac{\mathrm{d}^2}{\mathrm{d}^2\alpha}f_1^{(N)}(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}J_{11} + \frac{\mathrm{d}}{\mathrm{d}\alpha}J_{12} \leq 0. \tag{B.173}$$

Alternatively, one can use the log-sum inequality (see e.g. p.29 of [14]) to derive this concavity result (we checked that only for the concavity of $f_1(\alpha)$.

B.17 Proof of Lemma 42

The joint pmf of $\mathbf{S}$ and $\mathbf{Y}$ is given by $P(\mathbf{s},\mathbf{y}) = Q(\mathbf{s}) \cdot W(\mathbf{y}|\mathbf{s})$. From the Markovianity of $Q(.)$ we have for any $\ell \in \mathcal{I}_N$

$$Q(\mathbf{s}_\ell|\mathbf{s}_{-N}^{\ell-1}) = Q(s_\ell|s_{\ell-1}), \tag{B.174}$$

$$Q(\mathbf{s}) = Q(s_{\ell-1}, s_\ell) \cdot Q(\mathbf{s}_{-N}^{\ell-2}|s_{\ell-1}, s_\ell) \cdot Q(\mathbf{s}_{\ell+1}^{N}|s_{\ell-1}, s_\ell), \tag{B.175}$$

and from the hidden Markov property that

$$P(\mathbf{y}|\mathbf{s}_{-N}^{\ell-1}) = P(\mathbf{y}_{-N+1}^{\ell-1}|\mathbf{s}_{-N}^{\ell-1}) \cdot P(\mathbf{y}_{\ell}^{N}|s_{\ell-1}), \tag{B.176}$$
$$P(\mathbf{y}|\mathbf{s}_{-N}^{\ell}) = P(\mathbf{y}_{-N+1}^{\ell-1}|\mathbf{s}_{-N}^{\ell-1}) \cdot P(\mathbf{y}_{\ell}^{N}|s_{\ell}, s_{\ell-1}), \tag{B.177}$$

therefore

$$\begin{aligned}
V(s_\ell|\mathbf{s}_{-N}^{\ell-1}, \mathbf{y}) &= \frac{P(\mathbf{s}_{-N}^{\ell}, \mathbf{y})}{P(\mathbf{s}_{-N}^{\ell-1}, \mathbf{y})} && \text{(B.178)}\\
&= \frac{Q(\mathbf{s}_{-N}^{\ell}) \cdot P(\mathbf{y}|\mathbf{s}_{-N}^{\ell})}{Q(\mathbf{s}_{-N}^{\ell-1}) \cdot P(\mathbf{y}|\mathbf{s}_{-N}^{\ell-1})} && \text{(B.179)}\\
&= \frac{Q(\mathbf{s}_{-N}^{\ell-1}) \cdot Q(s_\ell|s_{\ell-1}) \cdot P(\mathbf{y}_{-N+1}^{\ell-1}|\mathbf{s}_{-N}^{\ell-1}) \cdot P(\mathbf{y}_{\ell}^{N}|s_\ell, s_{\ell-1})}{Q(\mathbf{s}_{-N}^{\ell-1}) \cdot P(\mathbf{y}_{-N+1}^{\ell-1}|s_{-N}^{\ell-1}) \cdot P(\mathbf{y}_{\ell}^{N}|s_{\ell-1})} && \text{(B.180)}\\
&= \frac{Q(s_\ell|s_{\ell-1}) \cdot P(\mathbf{y}_{\ell}^{N}|s_\ell, s_{\ell-1})}{P(\mathbf{y}_{\ell}^{N}|s_{\ell-1})} && \text{(B.181)}\\
&= \frac{Q(s_{\ell-1}, s_\ell) \cdot P(\mathbf{y}_{\ell}^{N}|s_\ell, s_{\ell-1})}{Q(s_{\ell-1}) \cdot P(\mathbf{y}_{\ell}^{N}|s_{\ell-1})} && \text{(B.182)}\\
&= \frac{P(s_{\ell-1}, s_\ell, \mathbf{y}_{\ell}^{N})}{P(s_{\ell-1}, \mathbf{y}_{\ell}^{N})} && \text{(B.183)}\\
&= V(s_\ell|s_{\ell-1}, \mathbf{y}_{\ell}^{N}) && \text{(B.184)}
\end{aligned}$$

Continuing from (B.182), we also have (using the equivalence $W(\mathbf{y}_{-N+1}^{\ell-1}|s_\ell, s_{\ell-1}) = W(\mathbf{y}_{-N+1}^{\ell-1}|s_{\ell-1}))$

$$\begin{aligned}
V(s_\ell|\mathbf{s}_{-N}^{\ell-1}, \mathbf{y}) &= \frac{Q(s_{\ell-1}, s_\ell) \cdot W(\mathbf{y}_{\ell}^{N}|s_\ell, s_{\ell-1})}{Q(s_{\ell-1}) \cdot W(\mathbf{y}_{\ell}^{N}|s_{\ell-1})} && \text{(B.185)}\\
&= \frac{Q(s_{\ell-1}, s_\ell) \cdot W(\mathbf{y}_{\ell}^{N}|s_\ell, s_{\ell-1}) \cdot W(\mathbf{y}_{-N+1}^{\ell-1}|s_\ell, s_{\ell-1})}{Q(s_{\ell-1}) \cdot W(\mathbf{y}_{\ell}^{N}|s_{\ell-1}) \cdot W(\mathbf{y}_{-N+1}^{\ell-1}|s_{\ell-1})} && \text{(B.186)}\\
&= \frac{P(s_{\ell-1}, s_\ell, \mathbf{y})}{P(s_{\ell-1}, \mathbf{y})} && \text{(B.187)}\\
&= V(s_\ell|s_{\ell-1}, \mathbf{y}). && \text{(B.188)}
\end{aligned}$$

This proves (B.1); proving (B.2) is along the same lines. With this, we can show (B.3)-(B.6)

$$\begin{aligned}
V(\mathbf{s}|\mathbf{y}) &= V(s_{\ell-1}, s_\ell|\mathbf{y}) \cdot V(\mathbf{s}_{-N}^{\ell-2}|s_{\ell-1}, s_\ell, \mathbf{y}) \cdot V(\mathbf{s}_{\ell+1}^{N}|\mathbf{s}_{-N}^{\ell}, \mathbf{y}) && \text{(B.189)}\\
&= V(s_{\ell-1}, s_\ell|\mathbf{y}) \cdot V(\mathbf{s}_{-N}^{\ell-2}|s_{\ell-1}, \mathbf{y}) \cdot V(\mathbf{s}_{\ell+1}^{N}|s_\ell, \mathbf{y}) && \text{(B.190)}\\
&= V(s_{\ell-1}, s_\ell|\mathbf{y}) \cdot V(\mathbf{s}_{-N}^{\ell-2}|s_{\ell-1}, \mathbf{y}_{-N+1}^{\ell-1}) \cdot V(\mathbf{s}_{\ell+1}^{N}|s_\ell, \mathbf{y}_{\ell+1}^{N}) && \text{(B.191)}\\
& && \text{(B.192)}
\end{aligned}$$

and

$$\begin{aligned}
V(\mathbf{s}|\mathbf{y}) &= V(s_\ell|\mathbf{y}) \cdot V(\mathbf{s}_{-N}^{\ell-1}|s_\ell, \mathbf{y}) \cdot V(\mathbf{s}_{\ell+1}^{N}|\mathbf{s}_{-N}^{\ell}, \mathbf{y}) && \text{(B.193)}\\
&= V(s_\ell|\mathbf{y}) \cdot V(\mathbf{s}_{-N}^{\ell-1}|s_\ell, \mathbf{y}) \cdot V(\mathbf{s}_{\ell+1}^{N}|s_\ell, \mathbf{y}) && \text{(B.194)}\\
&= V(s_\ell|\mathbf{y}) \cdot V(\mathbf{s}_{-N}^{\ell-1}|s_\ell, \mathbf{y}_{-N+1}^{\ell}) \cdot V(\mathbf{s}_{\ell+1}^{N}|s_\ell, \mathbf{y}_{\ell+1}^{N}). && \text{(B.195)}
\end{aligned}$$

B.18 Proof of Lemma 44

Remark: although the following derivation might look quite lengthy, the idea behind it is quite simple. To give it, we look at a simplified example. So, let $g(\alpha) \stackrel{\triangle}{=} g_1(\alpha) \cdot g_2(\alpha)/g_3(\alpha)$. Then

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} g(\alpha) = \frac{g_2(\alpha)}{g_3(\alpha)} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} g_1(\alpha) \right) + \frac{g_1(\alpha)}{g_3(\alpha)} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} g_2(\alpha) \right) - \frac{g_1(\alpha) g_2(\alpha)}{g_3^2(\alpha)} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} g_3(\alpha) \right) \tag{B.196}$$

$$= \frac{g(\alpha)}{g_1(\alpha)} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} g_1(\alpha) \right) + \frac{g(\alpha)}{g_2(\alpha)} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} g_2(\alpha) \right) - \frac{g(\alpha)}{g_3(\alpha)} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} g_3(\alpha) \right). \tag{B.197}$$

For proving the first part of the Lemma we start with $Q(\mathbf{s})$ as given in (3.7), i.e.,

$$Q(\mathbf{s}) = \frac{\prod_{\ell \in \mathcal{I}_N} Q_{s_{\ell-1}, s_\ell}}{\prod_{\ell \in \mathcal{I}'_N} \mu_{s_\ell}}, \tag{B.198}$$

or logarithmically,

$$\log(Q(\mathbf{s})) = \left(\sum_{\ell \in \mathcal{I}_N} \log\left(Q_{s_{\ell-1}, s_\ell}\right) \right) - \left(\sum_{\ell \in \mathcal{I}'_N} \log\left(\mu_{s_\ell}\right) \right) \tag{B.199}$$

$$= \left(\sum_{(i,j) \in \mathcal{A}} \sum_{\substack{\ell \in \mathcal{I}_N \\ s_{\ell-1}=i, s_\ell=j}} \log\left(Q_{s_{\ell-1}, s_\ell}\right) \right) - \left(\sum_{i \in \mathcal{S}} \sum_{\substack{\ell \in \mathcal{I}'_N \\ s_\ell = i}} \log\left(\mu_{s_\ell}\right) \right) \tag{B.200}$$

$$= \left(\sum_{(i,j) \in \mathcal{A}} \sum_{\substack{\ell \in \mathcal{I}_N \\ s_{\ell-1}=i, s_\ell=j}} \log\left(Q_{ij}\right) \right) - \left(\sum_{i \in \mathcal{S}} \sum_{\substack{\ell \in \mathcal{I}'_N \\ s_\ell = i}} \log\left(\mu_i\right) \right). \tag{B.201}$$

We have

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \log(Q(\mathbf{s})) = \left(\sum_{(i,j) \in \mathcal{A}} \sum_{\substack{\ell \in \mathcal{I}_N \\ s_{\ell-1}=i, s_\ell=j}} \frac{1}{Q_{ij}} Q_{ij}^{\alpha} \right) - \left(\sum_{i \in \mathcal{S}} \sum_{\substack{\ell \in \mathcal{I}'_N \\ s_\ell = i}} \frac{1}{\mu_i} \mu_i^{\alpha} \right) \tag{B.202}$$

$$= \left(\sum_{(i,j) \in \mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell \in \mathcal{I}_N \\ s_{\ell-1}=i, s_\ell=j}} \frac{1}{Q_{ij}} \right) - \left(\sum_{i \in \mathcal{S}} \mu_i^{\alpha} \sum_{\substack{\ell \in \mathcal{I}'_N \\ s_\ell = i}} \frac{1}{\mu_i} \right). \tag{B.203}$$

But from $\frac{\mathrm{d}}{\mathrm{d}\alpha} \log(Q(\mathbf{s})) = (\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}))/Q(\mathbf{s})$ it follows that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) = Q(\mathbf{s}) \cdot \frac{\mathrm{d}}{\mathrm{d}\alpha} \log(Q(\mathbf{s})) = \left(\sum_{(i,j) \in \mathcal{A}} Q_{ij}^{\alpha} \sum_{\substack{\ell \in \mathcal{I}_N \\ s_{\ell-1}=i, s_\ell=j}} \frac{Q(\mathbf{s})}{Q_{ij}} \right) - \left(\sum_{i \in \mathcal{S}} \mu_{s_\ell}^{\alpha} \sum_{\substack{\ell \in \mathcal{I}'_N \\ s_\ell = i}} \frac{Q(\mathbf{s})}{\mu_i} \right). \tag{B.204}$$

From $\sum_{i\in\mathcal{S}} \mu_i = \sum_{(i,j)\in\mathcal{A}} Q_{ij}$ and therefore $\sum_{i\in\mathcal{S}} \mu_i^\alpha = \sum_{(i,j)\in\mathcal{A}} Q_{ij}^\alpha$ it finally follows that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) = \left(\sum_{(i,j)\in\mathcal{A}} Q_{ij}^\alpha \sum_{\substack{\ell\in\mathcal{I}_N \\ s_{\ell-1}=i, s_\ell=j}} \frac{Q(\mathbf{s})}{Q_{ij}} \right) - \left(\sum_{(i,j)\in\mathcal{A}} Q_{ij}^\alpha \sum_{\substack{\ell\in\mathcal{I}_N' \\ s_\ell=i}} \frac{Q(\mathbf{s})}{\mu_i} \right). \tag{B.205}$$

The second result follows from

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \sum_{\mathbf{s}} Q(\mathbf{s}) \log Q(\mathbf{s}) = \sum_{\mathbf{s}} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) \log Q(\mathbf{s}) + \sum_{\mathbf{s}} Q(\mathbf{s}) \frac{1}{Q(\mathbf{s})} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) \tag{B.206}$$

$$= \sum_{\mathbf{s}} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) \log Q(\mathbf{s}) + \left(\sum_{\mathbf{s}} \frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) \tag{B.207}$$

$$= \sum_{\mathbf{s}} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) \log Q(\mathbf{s}) + \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} \underbrace{\sum_{\mathbf{s}} Q(\mathbf{s})}_{=1} \right) \tag{B.208}$$

$$= \sum_{\mathbf{s}} \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} Q(\mathbf{s}) \right) \log Q(\mathbf{s}). \tag{B.209}$$

Bibliography

[1] A. Kavčić, "On the capacity of Markov sources over noisy channels," in *Proc. IEEE GLOBECOM*, (San Antonio, TX, USA), pp. 2997–3001, Nov. 2001.

[2] D. Arnold and H.-A. Loeliger, "On the information rate of binary-input channels with memory," in *Proc. 2001 IEEE Int. Conf. on Communications*, (Helsinki, Finland), pp. 2692–2695, June 11–14 2001.

[3] D. Arnold, *Computing Information Rates of Finite-State Models with Application to Magnetic Recording.* PhD thesis, Swiss Federal Institute of Technology, ETH Zurich, 2002.

[4] V. Sharma and S. K. Singh, "Entropy and channel capacity in the regenerative setup with applications to markov channels," in *Proc. IEEE Intern. Symp. on Inform. Theory*, (Washington, D.C.), p. 283, June 24–29 2001.

[5] H. D. Pfister, J. B. Soriaga, and P. H. Siegel, "On the achievable information rates of finite-state ISI channels," in *Proc. IEEE GLOBECOM*, (San Antonio, TX), pp. 2992–2996, Nov. 2001.

[6] P. O. Vontobel and D. Arnold, "An upper bound on the capacity of channels with memory and constraint input," in *Proc. IEEE Inform. Theory Workshop*, (Cairns, Australia), pp. 147–149, Sept. 2-7 2001.

[7] R. E. Blahut, "Computation of channel capacity and rate distortion functions," *IEEE Trans. on Inform. Theory*, vol. IT–18, no. 4, pp. 460–473, 1972.

[8] S. Arimoto, "An algorithm for computing the capacity of arbitrary memoryless channels," *IEEE Trans. on Inform. Theory*, vol. IT–18, no. 1, pp. 14–20, 1972.

[9] J. A. O'Sullivan, "Alternating minimization algorithms: from Blahut-Arimoto to expectation-maximization," in *Codes, Curves, and Signals* (A. Vardy, ed.), pp. 173–192, Kluwer Academic Publisher, 1998.

[10] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[11] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. on Inform. Theory*, vol. IT–20, pp. 284–287, Mar. 1974.

[12] H.-A. Loeliger, "A posteriori probabilities and performance evaluation of trellis codes," in *Proc. IEEE Intern. Symp. on Inform. Theory*, (Trondheim, Norway), p. 335, June 27 –July 1 1994.

[13] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July/Oct. 1948.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[15] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. on Inform. Theory*, vol. IT–47, no. 2, pp. 498–519, 2001.

[16] G. D. Forney, Jr., "Codes on graphs: normal realizations," *IEEE Trans. on Inform. Theory*, vol. 47, no. 2, pp. 520–548, 2001.

[17] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis.* Cambridge: Cambridge University Press, 1994. Corrected reprint of the 1991 original.