Computation of the Rate-Distortion Function Relative to a Parametric Class of Reproductions

Yuval Kochman and Ram Zamir

Dpt. of EE-Systems, Tel Aviv University, Israel
yuvalko&zamir@eng.tau.ac.il

September 29, 2003

Abstract

The Blahut-Arimoto (BA) algorithm is a well known iterative procedure for computing the rate-distortion function. The Estimate-Maximize (EM) algorithm is a useful algorithm for finding the maximum likelihood parameter estimate from measurements. In this paper we consider the problem of computing the rate-distortion function relative to a parametric class of reproductions. We present a novel algorithm for this minimization problem, interchanging between stages of BA and EM. This algorithm converges to the rate-distortion function relative to the class of reproductions, whenever the parametric class is convex.

Keywords: Rate-distortion function, Blahut-Arimoto algorithm, maximum likelihood estimation, EM algorithm, alternating optimization, lossy coding, mismatched coding.

I Introduction

Given a memoryless source X with distribution P and some single-letter nonnegative distortion measure ρ, the rate-distortion function is defined by [2]:

$$\begin{aligned} R(P,d) &\stackrel{\Delta}{=} \min_{W:\rho(P,W)\leq d} I(P,W) \\ &= \min_{W:\rho(P,W)\leq d} D(P\circ W \| P\times [P\circ W]_y) \end{aligned} \tag{1}$$

where $I(\cdot)$ denotes mutual information, $\rho(P,W)=\sum_{x,y}P(x)W(y|x)\rho(x,y)$ is the average distortion induced by the input distribution P and transition distribution W, $P\circ W$ denotes a joint distribution induced by these two distributions: $[P\circ W](x,y)=P(x)W(y|x)$, $P\times Q$ denotes a product distribution: $[P\times Q](x,y)=P(x)Q(y)$, and $[\cdot]_y$ is a marginalization operator with respect to the reproduction distribution: $[S]_y(y)=\sum_x S(x,y)$. $D(\cdot\|\cdot)$ is the divergence, or relative entropy between two distributions [5].

If $W^*(P,d)$ is the transition distribution minimizing (1), then the optimal reproduction distribution $Q^*(P,d)$ is the induced marginal output distribution:

$$Q^*(P,d)=[P\circ W^*(P,d)]_y \quad , \tag{2}$$

is such that a random code generated i.i.d. $\sim Q^*(P,d)$ asymptotically achieves the rate given by this function [2]. If the random codebook is drawn using some other

*This work was supported in part by the United States - Israel Binational Science Fund, under grant 1998-309

reproduction distribution $Q \neq Q^*(P,d)$, occurs a situation known as *mismatched coding*. In recent years the subject of mismatched coding has received much attention (c.f. [12], [13], [14]). The optimal rate for memoryless P and reproduction Q is given by:

$$R(P,Q,d) = \min_{W:\rho(P,W)\leq d} D(P \circ W \| P \times Q) \ . \tag{3}$$

In a previous work [11] we considered the rate-distortion function relative to a class of reproductions $\mathcal{Q}$:

$$\begin{aligned} R(P,\mathcal{Q},d) &\triangleq \min_{Q\in\mathcal{Q}} R(P,Q,d) \\ &= \min_{Q\in\mathcal{Q}} \min_{W:\rho(P,W)\leq d} D(P \circ W \| P \times Q) \ . \end{aligned} \tag{4}$$

This function gives the optimal assymptotical rate of a random codebook if the reproduction distribution Q is restricted to be in the class $\mathcal{Q}$. It may be seen as the source coding quantity analogous to the constrained channel capacity of channel coding. This is of special practical interest when the class is a parametric class $\mathcal{Q}_\Theta$, since parametric classes of random codebooks may be constructed by a fixed random codebook followed by a parametric transformation [11]. A well known example for such a coding scheme is Code Excited Linear Prediction (CELP) used in speech coding [9]. We shall onwards restrict the discussion to parametric classes only, using the notation:

$$R(P,\Theta,d) \triangleq R(P,\mathcal{Q}_\Theta,d) = \min_{\theta\in\Theta} R(P,Q_\theta,d) \tag{5}$$

In general, minimization problems such as (1) and (5) can not be readily computed. For (1), the Blahut-Arimoto algorithm [3] gives an iterative way to compute any point on the R-d curve. Since the $R-d$ function is monotonous non-increasing and convex (as a function of d), it can be parametrically expressed using its slope[1]. Indeed, the BA algorithm assumes a Lagrange parameter s and iteratively creates a sequence of reproductions $Q^i, i = 1,2\ldots$ that converges to an optimal reproduction for any initial guess Q^0 bounded away from zero. Namely, for discrete, memoryless sources, the iteration:

$$\begin{aligned} W^{i+1}(y|x) &= \frac{Q^i(y)\exp(-s\rho(x,y))}{\sum_{y'} Q^i(y')\exp(-s\rho(x,y'))} \\ Q^{i+1}(y) &= \sum_x P(x)W^{i+1}(y|x), \quad i = 0,1,\ldots \end{aligned} \tag{6}$$

converges to the distribution $Q^*(P,d_s)$ minimizing (1) at the point of slope $-s$ of the $R(P,d)$ vs. d curve.

In this paper we present extensions of the BA algorithm for computing (5). Our approach is based upon the observation of [6] that the BA algorithm is a special case of alternating minimization between convex sets. In Section II we introduce our first algorithm, which uses at each iteration a projection from the reproduction computed by BA to the class $\mathcal{Q}_\Theta$. While in general this projection does not rely on the parametric representation of Q_Θ, in Section III we present our second algorithm which, inspired by the connection between this projection and maximum likelihood estimation, combines the BA algorithm with the EM algorithm [7]. In Section IV we address the issue of convergence rate of the algorithms. In Section V we discuss applicability of algorithms beyond the discrete, memoryless case. In Section VI we show an example of applying these algorithms to classes of memoryless mixture distributions. We conclude in Section VII by discussing similar algorithms for computation of channel capacity.

[1] If exist linear sections on the curve, they can be expressed by linear interpolation between their edges.

II Constrained BA: Three Stage Algorithm

The BA algorithm for computation of the $R-d$ function can be put in terms of alternating minimization by noting that the rate-distortion Lagrangian may be restated as a double minimization of a divergence (for $s > 0$):

$$\begin{aligned} G_s(P) &\stackrel{\Delta}{=} \min_d [R(P,d) + sd] \\ &= \min_Q \min_W D(P \circ W \| P \times Q \cdot \exp(-s\rho)) \ . \end{aligned} \tag{7}$$

Defining the sets: $\mathcal{A} = \{P \times Q \cdot \exp(-s\rho) : any\ Q\}$ and $\mathcal{B} = \{P \circ W : any\ W\}$, we see that each minimization corresponds to one of these sets:

$$G_s(P) = \min_{A \in \mathcal{A}} \min_{B \in \mathcal{B}} D(B\|A) \tag{8}$$

Note that the points in $\mathcal{A}$ are measures which are not probability distributions, yet divergence is defined the same way[2]. Since both sets are convex, an alternating minimization procedure, keeping at each time the chosen point in one of the sets fixed and minimizing the divergence relative to the other set, converges to the global optimum [6, Th. 3]. It can be shown, that the two steps of (6) are minimizations with respect to W and Q respectively, thus with respect to B and A.

The Lagrangian corresponding with the parametric reproduction in (5) is identical to (7), except that the minimization with respect to Q is confined to $Q \in \mathcal{Q}_\Theta$:

$$G_s(P,\Theta) = \min_{A \in \mathcal{A}_\Theta} \min_{B \in \mathcal{B}} D(B\|A) \tag{9}$$

where $\mathcal{A}_\Theta = \{P \times Q_\theta \cdot \exp(-s\rho) : \theta \in \Theta\}$. Thus the minimization with respect to W remains unchanged, while we need to find a new minimization for the closest element in the parametric class. Fortunately, this minimization can be broken into two stages [11, Th. 1], as demonstrated in Figure 1: First find the closest Q in the unconstrained set $\mathcal{A}$ as in the second stage of (6), and then find the distribution within $\mathcal{Q}_\Theta$ closest to that distribution in the divergence sense. All in all we have a three-step iterative procedure:

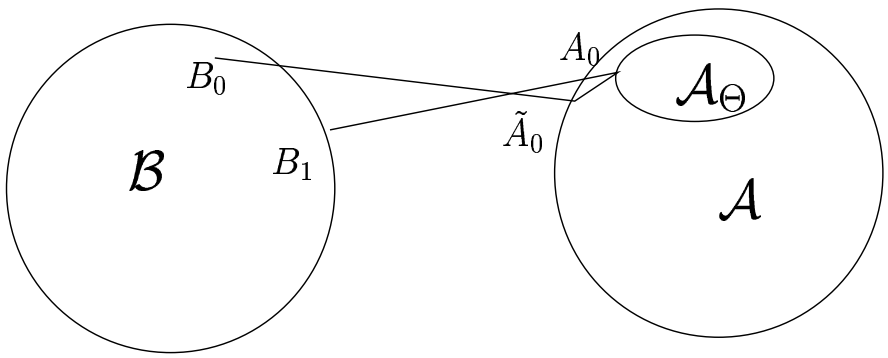

Figure 1: 3-stage Constrained Minimization

$$\begin{aligned} W^{i+1}(y|x) &= \frac{Q_\theta^i(y)\exp(-s\rho(x,y))}{\sum_{y'} Q^i(y')\exp(-s\rho(x,y'))} \\ \tilde{Q}^{i+1}(y) &= \sum_x P(x) W^{i+1}(y|x) \\ Q^{i+1} &= \arg\min_{Q \in \mathcal{Q}_\Theta} D(\tilde{Q}^{i+1}\|Q) \quad i = 0,1,2,\ldots. \end{aligned} \tag{10}$$

[2]Divergence between general measures (which do not sum up to one) may be negative. However, convexity holds, and theorems in [6] do not require measures to be distributions

By breaking the iteration into these steps, we have now isolated the effect of constraining the output distributions, and made it an additional stage to th BA algorithm. For some parametric classes this additional stage can be easily computed, while for others this is a non-trivial minimization problem. In the next section we use the parametric nature of Q_Θ in order to circumvent this difficulty for some parametric classes.

III Combining BA with EM: Four Stage Algorithm

First note that the last stage of (10) can be re-written as:

$$Q^{i+1} = \arg\max_{Q \in \mathcal{Q}_\Theta} E_{\tilde{Q}^{i+1}}\{\log Q\} \tag{11}$$

where

$$E_Q\{f(y)\} \triangleq \sum_y Q(y) f(y)$$

denotes expectation with respect to a distribution Q. This allows us to make a connection with estimation theory: Given a vector of measurements $y^n \triangleq y_1, \ldots, y_n$, and some parametric class of distributions $\mathcal{Q}_\Theta$, the maximum likelihood estimate for θ is given by:

$$\begin{aligned} \theta_{ML}(y^n) &\triangleq \arg\max_\theta Q_\theta(y^n) = \arg\max_\theta \log Q_\theta(y^n) \\ &= \arg\max_\theta E_{Q_{y^n}}\{\log Q_\theta\} \end{aligned} \tag{12}$$

where Q_{y^n} is the empirical distribution of the measurements y^n.

Comparing (11) with (12) we see that the divergence minimization in (10) is equivalent to maximum likelihood estimation of the parameter associated with a vector of measurements which have an empirical distribution $\tilde{Q}^{i+1}$.

From the computational point of view, this observation is very useful, since great efforts have been already put into the problem of maximum likelihood estimation. One of the approaches heavily investigated is of the EM algorithm [7], which has many applications in signal processing (see, c.f. [8]). This algorithm assumes that we have at hand measurements from a random variable Y having a joint distribution with some hidden variable Z, and thus is useful for cases where it is easier to estimate θ assuming a joint (Y, Z) distribution rather than directly from the empirical measure of y^n. Let the parametric set of joint (Y, Z) distributions be $\mathcal{T}_\Theta = \{T_\theta(y, z) : \theta \in \Theta\}$. Assume an initial guess θ^0 for the parameter. If $\mathcal{T}_\Theta$ is convex, then the EM iteration:

$$\begin{aligned} S^{i+1}(z|y) &= \frac{T^{\theta^i}(y, z)}{\sum_{z'} T_{\theta^i}(y, z')} \\ \theta^{i+1} &= \theta_{ML}(Q_y \circ S^{i+1}) \quad , i = 0, 1, \ldots \end{aligned} \tag{13}$$

converges to θ_{ML} [7].

In cases where the EM algorithm is useful, we now have a way for explicitly computing (10), though with many iterations of (13) for each iteration of (10)[3]. To further reduce the computation complexity we introduce the four stage algorithm, in which each iteration is composed of one iteration of the EM algorithm and one iteration of the BA algorithm: Start from an initial parameter θ^0 and iterate

$$W^{i+1}(y|x) = \frac{[T_{\theta^i}]_y(y) \exp(-s\rho(x, y))}{\sum_{y'} [T_{\theta^i}]_y(y') \exp(-s\rho(x, y'))}$$

[3]Theoretically, an infinite number of iterations is required. If we use some stopping condition, we will get a distribution that is near the ML one, and thus near the minimum divergence. As we show in the sequel, convergence is guarenteed regardless of the number of BA steps taken.

$$\begin{aligned}
\tilde{Q}^{i+1}(y) &= \sum_x P(x)W^{i+1}(y|x) \\
S^{i+1}(z|y) &= \frac{T_{\theta^i}(y,z)}{\sum_{z'} T_{\theta^i}(y,z')} \\
\theta^{i+1} &= \theta_{ML}(\tilde{Q}^{i+1} \circ S^{i+1}) \quad i = 0,1,2,\ldots
\end{aligned} \tag{14}$$

The following Theorem assures convergence of the algorithm. It is inspired by the fact that the EM algorithm is a special case of alternating minimization, just as the BA algorithm is [6]:

$$\theta_{ML}(y^n) = \arg\min_\theta \min_S D(Q_{y^n} \circ S \| T_\theta) \tag{15}$$

where we identify the two steps of (13) as minimization with respect to S and θ respectively. We define "super-sets" over the product alphabet of the source, the reproduction and the hidden variables, and show that algorithm (14) materializes alternating minimization between these sets.

Theorem 1 *If the set of distributions over (Y,Z) parametrized by Θ is convex, and $\mathcal{Q}_\Theta = \{[T_\theta]_y : \theta \in \Theta\}$, then for any initial guess θ^0 such that T^0_θ is bounded away from zero the iteration (14) converges to $\theta^* \stackrel{\Delta}{=} \arg\min_{\theta\in\Theta} R(P,Q_\theta,d)$ corresponding with optimum reproduction distribution $Q^*(P,\Theta,d)$ which achieves a point of slope $-s$ on the $R(P,\Theta,d)$ vs. d curve.*

For the proof, we need the following:

Lemma 1 *For any distributions $P(x)$, $W(y|x)$, $S(z|x,y)$ and $T(y,z) \in \mathcal{T}$, and a distance measure $\rho(x,y)$ we have:*

a. $\arg\min_{W,S} D(P \circ W \circ S \| P \times T \cdot \exp(-s\rho)) = \{W_1, S_1\}$, *where*

$$W_1 = \arg\min_W D(P \circ W \| P \times [T]_Y \cdot \exp(-s\rho)))$$

$$S_1 = [T]_{Z|Y}$$

b. $\arg\min_{T\in\mathcal{T}} D(P \circ W \circ S \| P \times T \cdot \exp(-s\rho)) = \arg\min_{T\in\mathcal{T}} D([P \circ W]_Y \circ S \| T)$

Proof:

$$\begin{aligned}
& D(P \circ W \circ S \| P \times T \cdot \exp(-s\rho)) \\
\geq\ & D(P \circ W \circ [S]_{Z|Y} \| P \times T \cdot \exp(-s\rho)) \\
=\ & \sum_x P(x) \sum_y W(y|x) \sum_z S(z|y) \log \frac{W(y|x)S(z|y)}{T(y,z)\exp(-s\rho(x,y))} \\
=\ & \sum_x P(x) \sum_y W(y|x) \{\log[W(y|x)\exp(s\rho(x,y))] + \sum_z S(z|y) \log \frac{S(z|y)}{T(y,z)}\}
\end{aligned} \tag{16}$$

where the inequality is due to the convexity of the divergence. The minimizations with respect to S and T are trivial since only the second term of (16) needs to be minimized. Then to see the result regarding W, we substitute the optimal S in (16):

$$D(P \circ W \circ S \| P \times T \cdot \exp(-s\rho))|_{S=S_1} = D(P \circ W \| P \times [T]_Y \cdot \exp(-s\rho)) \tag{17}$$

□

Proof of Theorem 1: We start by restating (14) as:

$$\begin{aligned} W^{i+1} &= \arg\min_{W} D(P \circ W \| P \times [T_{\Theta^i}]_Y \cdot \exp(-s\rho)) \\ S^{i+1} &= [T_{\Theta^i}]_{Z|Y} \\ \theta^{i+1} &= \arg\min_{\theta \in \Theta} D([P \circ W^{i+1}]_Y \circ S^{i+1} \| T_\Theta) \end{aligned} \tag{18}$$

where the first minimization corresponds to the BA minimization with respect to W, the second minimization is the EM minimization with respect to S, while the third minimization combines the BA minimization with respect to Q with the ML minimization with respect to Θ. Now we recognize that the minimizations (18) are actually the ones of Lemma 1, thus our iteration minimizes

$$D(P \circ W \circ S \| P \times T \cdot \exp(-s\rho))$$

in an alternating manner. Thus, if we define the set $\mathcal{A} = \{P \circ T \cdot \exp(-s\rho) \ : \ T \in \mathcal{T}_\Theta\}$ and the set $\mathcal{B}$ as all joint distributions over (X, Y, Z) with X-marginal P, we have that (14) is an alternating minimization between the two sets. For a convex $\mathcal{T}$, the sets $\mathcal{A}$ and $\mathcal{B}$ are convex in Q_Θ and (W, S) respectively, thus convergence is ensured by [6, Th. 3].

To see that this also solves the original minimization problem, suppose that we reached a minimizing triplet W^*, S^*, Θ^*. Substituting the expression for optimal S in (16), we get:

$$\min_{W,S,\theta} D(\mathcal{B} \| \mathcal{A}) = D(P \circ W^* \| P \times [T_{\theta^*}]_y \cdot \exp(-s\rho)) \quad .$$

This identity shows that a pair W, θ minimizes $D(\mathcal{B}\|\mathcal{A})$ if and only if it minimizes the rate-distortion Lagrangian, and that completes the proof □

Remark: We presented two algorithms, which may be seen as "one (BA) + many (EM)" and "one + one". It is only natural to ask which of them converges faster, and whether other combinations of BA and EM converge as well. It is easy to see, using Lemma 1, that any number of EM steps may be taken. Each additional step can only decrease the second term of (16), thus the sequence of errors is bounded from above by the sequence generated by the "one+one" algorithm. However, simulations show that one EM iteration is almost as good as many, thus the complexity of additional steps may not be worthwhile. The possibility of inserting additional BA steps, on the other hand, does not follow from this derivation, and on the contrary: We have counter examples where such steps prevent convergence to the global optimum.

IV An Upper bound on the Speed of Convergence

We are interested in the speed of convergence of the approximation error,

$$e_n \stackrel{\Delta}{=} R(P, Q_n, d) + s \cdot d - G_s(P) \quad . \tag{19}$$

For discrete memoryless sources, this error decays at least as inversely proportional to the number of iterations [4] [4] :

$$e_n \leq \frac{D(Q^* \| Q^0)}{n+1} \quad . \tag{20}$$

We will expand this bound to our algorithms.

[4] The original result is in terms of an initial guess for the transition distribution, but it is easy to translate it to terms of the marginal

Theorem 2 *For the iterations (10) and (14), convergence is at least inversely proportional to the number of iteration. Namely, (20) holds for (10), while for (14):*

$$e_n \leq \frac{D(T^*\|T^0)}{n+1} \quad ,$$

where $T^* = T(\Theta^*)$ is the joint distribution of the model used for EM, with the parameters of the distribution achieving $R(P,\Theta,d)$.

Proof: In both algorithms, we have alternating minimization of divergence between two sets. Denote the right hand and left hand sets by $\mathcal{A}$ and $\mathcal{B}$, the sequences generated by the iterations as A^n and B^n and the points of minimum divergence as A^* and B^*, respectively. Namely, for (10) we have $A^n = P \times Q^n \cdot \exp(-s\rho), B^n = P \circ W^n$ and for (14) we have $A^n = P \times T^n \cdot \exp(-s\rho),\ B^n = P \circ W^n \circ S^n$. In these terms, for both iterations we have:

$$e^n = D(B^{n+1}\|A^n) - D(B^*\|A^*) \leq D(B^n\|A^n) - D(B^*\|A^*) \tag{21}$$

For (10) this is straightforward, while for (14) it uses (17).

Now, by the "three point" and "four point" properties of [6, Th. 3]:

$$D(B^n\|A^n) - D(B^*\|A^*) \leq D(B^*\|A^n) - D(B^*\|A^{n+1}) \quad ,$$

thus combining with (21) and summing this inequality over iterations, we get:

$$\sum_{i=0}^{n} e^i \leq D(B^*\|A^0) - D(B^*\|A^{n+1})$$

and using the non-negativity of divergence and the monotonicity of the errors sequence e^i we can finally assert:

$$e^n \leq \frac{D(B^*\|A^0) - D(B^*|A^{n+1})}{n+1} \tag{22}$$

Now we go back to the explicit divergences: For (10) we have

$$D(B^*\|A^0) - D(B^*|A^{n+1}) = E_{P(X)W^*(Y|X)} \log \frac{Q^{n+1}(y)}{Q^0(y)} = E_{Q^*(Y)} \log \frac{Q^{n+1}(y)}{Q^0(y)} \leq D(Q^*\|Q^0)$$

and for (14) it can be shown that

$$D(B^*\|A^0) - D(B^*|A^{n+1}) \leq D(T^*\|T^0)$$

using the fact that $T^*(Y,Z) = [P(X) \circ W^*(Y|X)]_Y \circ S^*(Z|Y)$ $\square$

V Beyond the Discrete Memoryless Case

The extension to sources with continuous alphabet is straightforward. Basically, summations have to be changed by integration, and we have to talk about infimum rather than minimum. The results of [6] refer to the limit of the divergence sequence being the distance between sets. To guarantee finite distance, in the case where both source and parametric family have well defined pdfs, we substitute the requirement for the initial guess to have positive distribution everywhere by a requirement for positive density everywhere. Other cases, where exists a mixture of discrete and continuous distributions, require a more careful treatment.

Regarding sources with memory, $R(P,Q,d)$ is still the minimum coding rate as long as Q is stationary and ergodic [13]. There are two distinct cases here: If the reproduction class is memoryless, then the optimal solution is just the same as the one for a memoryless source with the same marginal ([13], [11]), thus algorithms may be applied to this marginal. If we allow reproductions with memory, algorithms have to be applied to multi-dimensional distributions. This is not always feasible, but at least for Markovian reproduction classes (such as a Gaussian-AR model we used in [11]) it is possible to look at finite-dimensional distributions.

VI Example: Memoryless Mixture Reproductions

In this example we consider classes of distributions composed of known memoryless distributions with weights defined by the parameter vector:

$$Q_Y(y) = \sum_{m=0}^{M-1} \theta_m Q_m(y) \ , \ \sum_{m=0}^{M-1} \theta_m = 1 \tag{23}$$

Maximum likelihood estimation for such a model is not immediate. An iterative solution was suggested in [10], and later identified to be a special case of the EM algorithm. For an empirical data distribution Q_y and an initial guess θ^0, EM iteration n would be:

$$\begin{aligned} C_m &= \sum_y Q_y(y) \frac{\theta_m^n Q_m(y)}{\sum_{m'} \theta_{m'}^n Q_{m'}(y)} \\ \theta_m^{*\ n+1} &= \frac{C_m}{\sum_{m'} C_{m'}} \quad m = 0, 1, \ldots, M-1 \end{aligned} \tag{24}$$

The example we have chosen is of a Gaussian mixture source, composed of three components of variance 0.1 each, centered around $(-1, 0, 1)$ with weights $(0.5, 0.25, 0.25)$ respectively. Reconstruction is by a Gaussian mixture of two components of variance 0.05 each, centered around $(-1, 1)$. Figure 2 shows the functions $R(P, d)$ and $R(P, \Theta, d)$ for this case. As expected, the $R - d$ function relative to the parametric class is always higher.

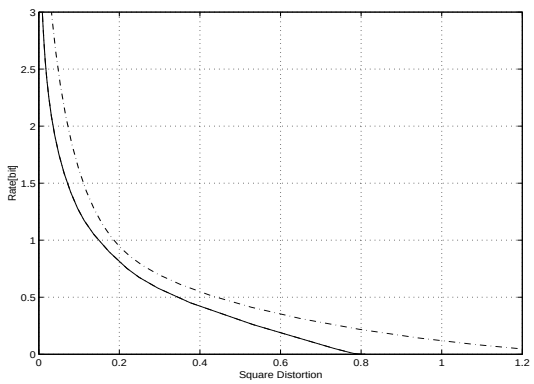

Figure 2: Constrained vs. Non-constrained $R - d$ function

Next we show the rate of convergence of our two algorithms. For this, we chose the point of slope (-1) on the $R(P, \Theta, d)$ curve. As shown in Figure 3, convergence is almost the same for both algorithms, in terms of iterations, but recall that while for the first algorithm each iteration contains multiple EM steps, for the second one it contains one such step only. This implies that indeed it is plausible to use the interleaving between AM and EM.

It is interesting to see, that both algorithms display exponential decay of the error. While this is a well known property of EM (see [7]), we did not find a stronger claim regarding BA than the one in [4]. Since for many unconstrained cases (c.f. finite alphabet, memoryless sources) it is possible to represent BA as a parametric minimization problem, it should be interesting to find general conditions for exponential convergence.

VII Discussion: Constrained Channel Capacity

The variant of the BA algorithm dealing with the computation of Channel capacity is perhaps better known than the one dealing with $R - d$ calculation. This variant of the

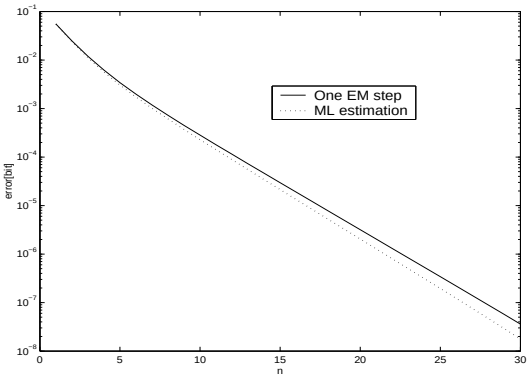

Figure 3: Convergence of proposed algorithms

algorithm was first presented in [1]: If the channel conditional distribution is $W(y|x)$, choose P^0, then iterate

$$\begin{aligned} \Phi^{i+1}(x|y) &= \frac{P^i(x)W(y|x)}{\sum_{x'} P^i(x')(W(y|x'))} \\ P^{i+1}(x) &= \frac{\prod_y((\Phi^{i+1}(x|y))^{W(y|x)})}{\sum_{x'}\prod_y((\Phi^{i+1}(x'|y))^{W(y|x')})} \quad i = 0, 1, 2, \ldots \end{aligned} \tag{25}$$

This variant was also shown in [6] to be a special case of alternating minimization of the divergence between convex sets. It can be shown that:

$$C(W) = -\min_{\Phi(x|y)} \min_{P(x)} D(P \circ W \| \Phi(x|y)W(y|x)) \tag{26}$$

where again, like in the R-d case, the elements on the right hand side of the divergence are general positive measures. If we define the sets $\mathcal{A} = \{P \circ W : \ any\ P\}$ and $\mathcal{B} = \{\Phi(x|y)W(y|x) : \ any\ \Phi(x|y)\}$, the two stages of (25) minimize the distance relative to each set.

The problem of finding capacity and optimal input distribution when these distributions are constrained, was already addressed in [3]. There, the capacity at expense E is defined as:

$$C(E) = \max_{P \in P(E)} I(P, W), \ P_E = \{P : \sum_x P(x)e(x) \leq E\} \tag{27}$$

For this average cost constraint on the input distribution, [3] gives an algorithm minimizing the information-expense Lagrangian, very similar to the minimization of the R-d Lagrangian discussed before, which is another special case of the alternating minimization algorithms.

We suggest to attack a more general problem, the computation of the channel capacity relative to any convex parametric constraint. By explicitly expending the divergence in (26), one can see that:

$$D(P \circ W \| \Phi^i(x|y)W(y|x)) = D(P\|P^i) + K(\Phi^i, W) \tag{28}$$

thus, as we did in the three-stage algorithm for R-d, we can apply the constraint by an additional stage:

$$\begin{aligned} \Phi^{i+1}(x|y) &= \frac{P^i_\theta(x)W(y|x)}{\sum_{x'} P^i_\theta(x')(W(y|x'))} \\ \tilde{P}^{i+1}(x) &= \frac{\prod_y((\Phi^{i+1}(x|y))^{W(y|x)})}{\sum_{x'}\prod_y((\Phi^{i+1}(x'|y))^{W(y|x')})} \\ \theta^{i+1} &= \arg\min_{\theta \in \Theta} D(P_\theta \| \tilde{P}^{i+1}) \end{aligned} \tag{29}$$

Efficient ways to compute the minimizer of the third stage are still to be investigated. If exists an "EM-like" way to minimize this divergence, it may also be possible to interchange between these iterations and the original BA iterations, as we did for the R-d computation.

References

[1] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Information Theory*, 18:14–20, Jan. 1972.

[2] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Prentice-Hall, Englewood Cliffs, NJ, 1971.

[3] R. E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Information Theory*, IT-18:460–473, 1972.

[4] P. Boukris. An upper bound on the speed of convergence of the Blahut algorithm for computing rate-distortion functions. *IEEE Trans. Information Theory*, pages 708–709, Sept. 1973.

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

[6] I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and decisions*, Supplement issue No. 1:205–237, 1984.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Ser. 39:1–38, 1977.

[8] M. Feder, A. V. Oppenheim, and E. Weinstein. Maximum Likelihood Noise Cancellation Using the EM Algorithm. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-37:204–216, Feb. 1989.

[9] G.D.Gibson, T.Berger, T.Lookabaugh, D.Lindbergh, and R.L.Baker. *Digital Compression for Multimedia: Principles and Standards.* Morgan Kaufmann Pub., San Fansisco, 1998.

[10] H. Hartley. Maximum likelihood estimation from incomplete data. *Biometrics*, 14:174–194, 1958.

[11] Y. Kochman and R. Zamir. Adaptive parametric vector quantization by natural type selection. In *Proc. of the Data Comp. Conf.*, pages 392–401, April 2002.

[12] A. Lapidoth. On the role of missmatch in rate-distortion theory. *IEEE Trans. Information Theory*, 43:38–47, Jan. 1997.

[13] E.H. Yang and J. Kieffer. On the performance of data compression algorithms based upon string matching. *IEEE Trans. Information Theory*, IT-44:47–65, Jan. 1998.

[14] R. Zamir and K. Rose. Natural type selection in addaptive lossy compression. *IEEE Trans. Information Theory*, 47:99–111, Jan. 2001.