Information Geometric Formulation and Interpretation of Accelerated Blahut-Arimoto-Type Algorithms

Gerald Matz* and Pierre Duhamel

Laboratoire des Signaux et Systèmes, Ecole Supérieure d'Electricité
3 Rue Joliot-Curie, F-91190 Gif-sur-Yvette, FRANCE
phone: +33 1 69 85 17 57, fax: +33 1 69 85 17 65, email: {gerald.matz,pierre.duhamel}@lss.supelec.fr

Abstract — **We propose two related iterative algorithms for computing the capacity of discrete memoryless channels. The celebrated Blahut-Arimoto algorithm is a special case of our framework. The formulation of these algorithms is based on the natural gradient and proximal point methods. We also provide interpretations in terms of notions from information geometry. A theoretical convergence analysis and simulation results demonstrate that our new algorithms have the potential to significantly outperform the Blahut-Arimoto algorithm in terms of convergence speed.**

I. Introduction

It is now exactly 30 years that R. Blahut and S. Arimoto both received the Information Theory Paper Award for their Jan. 1972 Transactions Papers on how to numerically compute channel capacity and rate-distortion functions [3, 4]. In [9], an information geometric interpretation of the Blahut-Arimoto (BA) algorithm in terms of alternating information projections was provided. Since then, several extensions of BA to other types of channels have been proposed (e.g. [10, 13, 15–17]).

Our contributions regarding capacity computation for discrete memoryless channels (DMCs) in this paper are:

- Capacity computation is shown to be equivalent to an information geometric "equidivergence" game (Section III).
- We propose a natural gradient (NG) algorithm [1] and an accelerated BA algorithm for capacity computation (Section IV). We demonstrate that close to the optimum, the accelerated BA and NG algorithms are approximately equivalent.
- We rephrase the accelerated BA and NG algorithms as proximal point methods that respectively use Kullback-Leibler divergence and chi-square divergence between the iterates as penalty terms (Section V).
- We provide a convergence analysis of the accelerated BA algorithm which roughly also characterizes the convergence of the NG algorithm (Section VI).
- Numerical experiments confirm our theoretical results and show that our novel algorithms converge significantly faster than the BA algorithm (Section VII).

*On leave from Institute of Communications and Radio-Frequency Engineering, Vienna University of Technology. Funded by FWF Grant J2302.

The relevant information geometric facts and notions used in the paper can be found in [2, 8].

After formulating the accelerated BA and NG algorithms, we discovered close relations to [5, 12] which consider modifications of the EM algorithm for ML estimation in general and mixture estimation in particular. In fact, some of our convergence results are inspired by [5].

II. Background

Consider a DMC with input symbol X taken from the size $M+1$ input alphabet $\{x_0, \ldots, x_M\}$, output symbol Y in the size $N+1$ alphabet $\{y_0, \ldots, y_N\}$, and transition probabilities $Q_{i|j} = \Pr(Y = y_i | X = x_j)$. We define the $(N+1) \times (M+1)$ channel matrix $\mathbf{Q}$ as $[\mathbf{Q}]_{ij} = Q_{i|j}$. The distribution of the input and output symbol are characterized, respectively, by the probability vectors $\mathbf{p} = [p_0 \ldots p_M]^T$ and $\mathbf{q} = [q_0 \ldots q_M]^T = \mathbf{Q}\mathbf{p}$ with $p_j = \Pr(X = x_j)$ and $q_i = \Pr(Y = y_i) = \sum_{j=0}^{M} Q_{i|j}\, p_j$. The mutual information of X and Y equals [6, 11]

$$I(\mathbf{p}) = \sum_{j=0}^{M} \sum_{i=0}^{N} p_j Q_{i|j} \log \frac{Q_{i|j}}{q_i} = \sum_{j=0}^{M} p_j D(\mathbf{Q}_j \| \mathbf{q})$$

where $\mathbf{Q}_j$ denotes the jth column of $\mathbf{Q}$ and we used the Kullback-Leibler divergence (KLD) [6], defined as

$$D(\mathbf{p} \| \mathbf{p}') = \sum_j p_j \log \frac{p_j}{p'_j}\,.$$

The capacity of the channel equals $C = \max_{\mathbf{p}} I(\mathbf{p})$ (this is a maximization of a continuous concave function over a closed convex set). The Kuhn-Tucker conditions for an optimum (i.e., capacity-achieving) input distribution $\mathbf{p}^*$ are [11]

$$D(\mathbf{Q}_j \| \mathbf{Q}\mathbf{p}^*) = C, \quad p_j^* > 0, \tag{1a}$$

$$D(\mathbf{Q}_j \| \mathbf{Q}\mathbf{p}^*) < C, \quad p_j^* = 0. \tag{1b}$$

We note that for any input distribution $\mathbf{p}$ we have [4, 11]

$$\sum_{j=0}^{M} p_j D(\mathbf{Q}_j \| \mathbf{Q}\mathbf{p}) \;\le\; C \;\le\; \max_j D(\mathbf{Q}_j \| \mathbf{Q}\mathbf{p})\,. \tag{2}$$

These inequalities become equalities in the case of a capacity-achieving input distribution and can be used as a termination criterion for all the iterative algorithms below.

The key for the BA algorithm [3, 4] is the observation that $C = \max_{\mathbf{p}} \max_{\mathbf{P}} J(\mathbf{p}, \mathbf{P})$ where

$$J(\mathbf{p}, \mathbf{P}) = \sum_{j=0}^{M} \sum_{i=0}^{N} p_j Q_{i|j} \log \frac{P_{j|i}}{p_i}$$

and $\mathbf{P}$ is an arbitrary $(M+1) \times (N+1)$ transition probability matrix with entries $[\mathbf{P}]_{ji} = P_{j|i}$ (in fact, $I(\mathbf{p}) = \max_{\mathbf{P}} J(\mathbf{p}, \mathbf{P})$). Starting from an initial guess $\mathbf{p}^0$, the BA algorithm iteratively computes

$$\mathbf{P}^{k+1} = \arg\max_{\mathbf{P}} J(\mathbf{p}^k, \mathbf{P}), \qquad \mathbf{p}^{k+1} = \arg\max_{\mathbf{p}} J(\mathbf{p}, \mathbf{P}^{k+1}). \tag{3}$$

In [9], these maximizations were re-interpreted as alternating projections based on KLD minimizations. Combining both maximizations into one step yields the multiplicative BA update

$$p_j^{k+1} = p_j^k \frac{\exp(D_j^k)}{\sum_{j=0}^{M} p_j^k \exp(D_j^k)}. \tag{4}$$

Here, $D_j^k \triangleq D(\mathbf{Q}_j \| \mathbf{q}^k)$ with $\mathbf{q}^k = \mathbf{Q}\mathbf{p}^k$.

III. An Equidivergence Game

Based on the arguments below, capacity computation is equivalent in information geometric terms to the following "equidivergence" game (for simplicity of exposition, we restrict to the case $p_j^* > 0, j = 0, \ldots, M$). Consider the set of length-$N+1$ probability vectors $\mathbf{q}$ that have the same KLD to all columns of $\mathbf{Q}$:

$$\mathcal{Q} = \{\mathbf{q} : D(\mathbf{Q}_0 \| \mathbf{q}) = D(\mathbf{Q}_1 \| \mathbf{q}) = \cdots = D(\mathbf{Q}_N \| \mathbf{q})\}. \tag{5}$$

It can be shown that this is a log-linear (exponential) [2, 8] family of probability vectors. Hence, the reverse I-projection [2, 8] of the jth column of $\mathbf{Q}$ onto $\mathcal{Q}$, defined as

$$\mathbf{q}^* = \arg\min_{\mathbf{q} \in \mathcal{Q}} D(\mathbf{Q}_j \| \mathbf{q}), \tag{6}$$

belongs to the linear (mixture) family

$$\mathcal{L} = \left\{ \mathbf{q} : \mathbf{q} = \mathbf{Q}\mathbf{p} = \sum_{j=0}^{N} \mathbf{Q}_j p_j, \quad \sum_{j=0}^{N} p_j = 1 \right\},$$

which is dual to $\mathcal{Q}$. Using the compensation identity

$$\sum_{j=0}^{N} p_j D(\mathbf{Q}_j \| \mathbf{Q}\mathbf{p}) = \sum_{j=0}^{N} p_j D(\mathbf{Q}_j \| \mathbf{q}) - D(\mathbf{Q}\mathbf{p} \| \mathbf{q}),$$

with $\mathbf{q} = \mathbf{q}^* \in \mathcal{Q}$, it follows that $\sum_{j=0}^{N} p_j D(\mathbf{Q}_j \| \mathbf{Q}\mathbf{p}) \leq D(\mathbf{Q}_j \| \mathbf{q}^*)$ (recall (5)) with equality iff $\mathbf{Q}\mathbf{p} = \mathbf{q}^*$. Thus,

$$\mathbf{q}^* = \mathbf{Q}\mathbf{p}^* \quad \text{with} \quad \mathbf{p}^* = \arg\max_{\mathbf{p}} \sum_{j=0}^{N} p_j D(\mathbf{Q}_j \| \mathbf{Q}\mathbf{p}).$$

We conclude that the equidivergence game (6) is equivalent to capacity computation.

From an algorithmic point of view, the equidivergence game means that given a current guess $\mathbf{p}^k$, we should check the KLDs $D_j^k = D(\mathbf{Q}_j \| \mathbf{Q}\mathbf{p}^k)$ and move the output distribution closer to those $\mathbf{Q}_j$ for which D_j^k is large. This can be achieved by increasing the respective weights p_j^k, consistent with the BA recursion (4) that increases (decreases) those input probabilities for which $\exp(D_j^k)$ is above (below) the average $\sum_{j=0}^{M} p_j \exp(D_j^k)$.

IV. Two New Algorithms

A. Natural Gradient Algorithm. We next exploit the fact that the input probability vectors $\mathbf{p}$ constitute an M-dimensional Riemannian manifold to propose a novel algorithm for capacity computation that is based on the *natural gradient* (NG) [1]. To this end, we describe the input probability vectors in terms of their M dual (or expectation) parameters $\eta_j = p_j$, $j = 1, \ldots, M$ (note that $p_0 = 1 - \sum_{j=1}^{M} \eta_j$). In terms of these parameters, the score function $\bar{I}(\boldsymbol{\eta}) = I(\mathbf{p})$ reads

$$\bar{I}(\boldsymbol{\eta}) = \sum_{j=1}^{M} \sum_{i=0}^{N} \eta_j Q_{i|j} \log \frac{Q_{i|j}}{q_i} + \left[1 - \sum_{j=1}^{M} \eta_j\right] \sum_{i=0}^{N} Q_{i|0} \log \frac{Q_{i|0}}{q_i}$$

with $q_i = \sum_{j=1}^{M} Q_{i|j} \eta_j + Q_{i|0}\left[1 - \sum_{j=1}^{M} \eta_j\right]$. To climb the peak of this score function, we propose a NG ascent algorithm with the parameter updates

$$\boldsymbol{\eta}^{k+1} = \boldsymbol{\eta}^k + \mu_k \widetilde{\nabla} \bar{I}(\boldsymbol{\eta}^k). \tag{7}$$

Here, μ_k is a step-size parameter and $\widetilde{\nabla} \bar{I}(\boldsymbol{\eta})$ is the NG of $\bar{I}(\boldsymbol{\eta})$ obtained by pre-multiplying the ordinary gradient with the inverse of the Riemannian metric $\mathbf{G}(\boldsymbol{\eta})$ (which here equals the Fisher information matrix of the parameters $\boldsymbol{\eta}$ [2]):

$$\widetilde{\nabla} \bar{I}(\boldsymbol{\eta}) = \mathbf{G}^{-1}(\boldsymbol{\eta}) \nabla \bar{I}(\boldsymbol{\eta}), \qquad \nabla \bar{I}(\boldsymbol{\eta}) = \left[\frac{\partial \bar{I}(\boldsymbol{\eta})}{\partial \eta_1} \cdots \frac{\partial \bar{I}(\boldsymbol{\eta})}{\partial \eta_M}\right]^T.$$

Computing $\mathbf{G}(\boldsymbol{\eta}) = \operatorname{diag}\{\boldsymbol{\eta}\} - \boldsymbol{\eta}\boldsymbol{\eta}^T$ and using the fact that the jth component of $\nabla \bar{I}(\boldsymbol{\eta})$ equals $D(\mathbf{Q}_j \| \mathbf{q}) - D(\mathbf{Q}_0 \| \mathbf{q})$, the jth component of the NG is obtained as

$$\left[\widetilde{\nabla} \bar{I}(\boldsymbol{\eta})\right]_j = \left[D(\mathbf{Q}_j \| \mathbf{q}) - \bar{I}(\boldsymbol{\eta})\right] \eta_j. \tag{8}$$

Note that the NG points towards the directions for which the KLD $D(\mathbf{Q}_j \| \mathbf{q})$ is large. Plugging (8) into (7), the NG recursions for the input probabilities can be shown to be

$$p_j^{k+1} = p_j^k \left[1 + \mu_k (D_j^k - I^k)\right]. \tag{9}$$

Here, we used the short-hand notation $I^k \triangleq I(\mathbf{p}^k)$ for the current estimate (actually, lower bound) of capacity. Due to $\sum_{j=0}^{M} p_j^k D_j^k = I^k$, (9) guarantees $\sum_{j=0}^{M} p_j^{k+1} = 1$. Non-negativity of p_j^{k+1} is ensured for $\mu_k \leq -\frac{1}{\min_j D_j^k - I^k}$. Note that like the BA recursion, (9) amounts to a multiplicative update. However, the computational complexity of the NG update is less than that of the BA update since the exponentiation is avoided. Furthermore, (9) is consistent with the information geometric equidivergence interpretation of capacity computation.

We observed that the NG algorithm can outperform the BA algorithm significantly in terms of convergence speed. The NG algorithm can be shown to replace the second maximization in (3) with a NG ascent step. While this step misses the local maximum along the $\mathbf{p}$-axis, it allows to better approach the global maximum, i.e., the NG algorithm basically avoids traversing back and forth a ridge of $J(\mathbf{p}, \mathbf{P})$.

B. Accelerated Blahut-Arimoto Algorithm. In our simulations, we observed similar convergence properties of the BA

and NG algorithms for $\mu = 1$. To explain why this is the case, we divide numerator and denominator of the BA update (4) by $\exp(I^k)$ and then use a first-order Taylor series approximation of the exponentials:

$$p_j^{k+1} = p_j^k \frac{\exp(D_j^k - I^k)}{\sum_{j=0}^M p_j^k \exp(D_j^k - I^k)} \approx p_j^k \left[1 + (D_j^k - I^k)\right].$$

The right-hand side is the NG recursion (9) for $\mu_k = 1$. The Taylor series approximation is accurate for $D_j^k - I^k \approx 0$, i.e., when equidivergence is almost achieved which will be true in the vicinity of the optimum solution. The above approximate equivalence motivates the *ad hoc* formulation of a generalized BA algorithm:

$$p_j^{k+1} = p_j^k \frac{\exp(\mu_k D_j^k)}{\sum_{j=0}^M p_j^k \exp(\mu_k D_j^k)}. \tag{10}$$

Using the same arguments as before, this algorithm can be shown to be asymptotically equivalent to the NG algorithm. In fact, using the same step-size $\mu_k = \mu$, both algorithms feature the same convergence speed which is μ times faster than that of the ordinary BA algorithm. For that reason, we refer to (10) as *accelerated BA* algorithm. The convergence behavior of the accelerated BA and NG algorithms and the choice of μ_k are discussed further in Sections V and VI.

C. Interpretation via e- and m-Geodesics. Next, we briefly discuss the information geometric significance of the accelerated BA and NG algorithms. Assume that the current guess for the optimum input and output probabilities are $\mathbf{p}$ and $\mathbf{q} = \mathbf{Q}\mathbf{p}$. Let $\mathbf{p}_{\mathrm{BA}}$ and $\mathbf{p}_{\mathrm{NG}}$ be the probabilities obtained by applying to $\mathbf{p}$ the BA and NG updates (10) and (9) with $\mu_k = 1$ and $\mu_k = 1/I^k$, respectively. It is then easily verified that the accelerated BA and NG updates for general μ_k can be written as

$$\mathbf{p}_{\mathrm{BA}}(\mu_k) = c(\mu_k)\, \mathbf{p}^{1-\mu_k}\, \mathbf{p}_{\mathrm{BA}}^{\mu_k},$$
$$\mathbf{p}_{\mathrm{NG}}(\mu_k) = (1 - \mu_k I_k)\, \mathbf{p} + \mu_k I_k\, \mathbf{p}_{\mathrm{NG}},$$

where $c(\mu_k)$ is a normalization constant. (Note that $\mathbf{p}_{\mathrm{BA}}(0) = \mathbf{p}_{\mathrm{NG}}(0) = \mathbf{p}$.) Hence, the probability vectors $\mathbf{p}_{\mathrm{BA}}(\mu_k)$ constitute an exponential (log-linear) family, parameterized by μ_k, that corresponds to the e-geodesic [2] connecting $\mathbf{p}$ and $\mathbf{p}_{\mathrm{BA}}$. In contrast, the $\mathbf{p}_{\mathrm{NG}}(\mu_k)$ constitute the mixture (linear) family, again parameterized by μ_k, that corresponds to the m-geodesic [2] connecting $\mathbf{p}$ and $\mathbf{p}_{\mathrm{NG}}$. For $|\mu_k(D_j^k - I^k)| \ll 1$, these two geodesics virtually coincide in the vicinity of $\mathbf{p}$.

V. Proximal Point Reformulation

We previously provided some information geometric insights regarding the accelerated BA and NG algorithms and investigated their asymptotic equivalence. In this section, we demonstrate that in fact both algorithms can be derived within a common framework that also provides an *a posteriori* justification for the *ad hoc* definition (10) of the accelerated BA recursions.

A. Accelerated BA Algorithm. The key observation is obtained by a re-examination of the alternating maximizations (3) underlying the BA algorithm. By plugging the explicit solution $P_{j|i}^{k+1} = Q_{i|j}\, p_j^k / \sum_{j'=0}^M Q_{i|j'}\, p_{j'}^k$ of the first maximization into the second maximization problem, we obtain

$$\mathbf{p}^{k+1} = \arg\max_{\mathbf{p}} \sum_{j=0}^M \sum_{i=0}^N p_j Q_{i|j} \log \frac{P_{j|i}^{k+1}}{p_j}$$
$$= \arg\max_{\mathbf{p}} \left\{ \sum_{j=0}^M p_j\, D_j^k - D(\mathbf{p}\|\mathbf{p}^k) \right\} \tag{11}$$

(recall $D_j^k = D(\mathbf{Q}_j\|\mathbf{Q}\mathbf{p}^k)$). (11) can be interpreted as a maximization of $\sum_{j=0}^M p_j\, D_j^k$ with a penalty term $D(\mathbf{p}\|\mathbf{p}^k)$ that ensures that the update $\mathbf{p}^{k+1}$ remains in the vicinity of $\mathbf{p}^k$. Algorithms of this type are known as *proximal point methods* [5] since they force the update to stay in the proximity of the current guess. In our case this is reasonable since the first term in (11), $\sum_{j=0}^M p_j\, D_j^k$, can be viewed as an approximation of the true score function $I(\mathbf{p})$ obtained by replacing the KLDs $D(\mathbf{Q}_j\|\mathbf{Q}\mathbf{p})$ with $D_j^k = D(\mathbf{Q}_j\|\mathbf{Q}\mathbf{p}^k)$. The penalty term $D(\mathbf{p}\|\mathbf{p}^k)$ in (11) ensures that the maximization is restricted to a neighborhood of $\mathbf{p}^k$ for which the approximation $D(\mathbf{Q}_j\|\mathbf{Q}\mathbf{p}) \approx D(\mathbf{Q}_j\|\mathbf{Q}\mathbf{p}^k)$ is accurate. In fact, by adding in (11) the quantity $I(\mathbf{p}^k) - \sum_{j=0}^M p_j^k\, D_j^k$ (which is independent of $\mathbf{p}$), we obtain

$$\mathbf{p}^{k+1} = \arg\max_{\mathbf{p}} \left\{ \tilde{I}^k(\mathbf{p}) - D(\mathbf{p}\|\mathbf{p}^k) \right\}, \tag{12}$$

where $\tilde{I}^k(\mathbf{p}) = I(\mathbf{p}^k) + \sum_{j=0}^M (p_j - p_j^k)\, D_j^k$ can be shown to be a first-order Taylor series approximation of $I(\mathbf{p})$ about $\mathbf{p}^k$ (the jth component of the gradient of $I(\mathbf{p})$ at $\mathbf{p}^k$ equals $D_j^k - 1$). Hence, the BA algorithm can be viewed as proximal point method maximizing the first-order Taylor series approximation of $I(\mathbf{p})$ with a proximity penalty expressed by $D(\mathbf{p}\|\mathbf{p}^k)$.

It is now natural to modify (12) by emphasizing/attenuating the penalty term via a weighting, i.e.,

$$\mathbf{p}^{k+1} = \arg\max_{\mathbf{p}} \left\{ \tilde{I}^k(\mathbf{p}) - \gamma_k\, D(\mathbf{p}\|\mathbf{p}^k) \right\}. \tag{13}$$

The idea is that close to the optimal solution the gradient of $I(\mathbf{p})$ will be small and thus the proximity constraint can be gradually relaxed by decreasing γ_k. It is straightforward to solve this problem using Lagrange multiplier techniques and it turns out that the optimum solution is given by (10) with $\mu_k = 1/\gamma_k$. We conclude that the accelerated BA update can be viewed as proximal point algorithm with weighted proximity penalty.

B. NG Algorithm. We next demonstrate that like the accelerated BA algorithm our NG algorithm can be viewed as proximal point method. The modification that is required pertains to the penalty term which in the accelerated BA algorithm is formulated in terms of the KLD. Obviously, there exist countless other distance functions that can be used to force the update to the vicinity of the current guess. Since we are iterating on the manifold of probability vectors, Euclidean distance is not a reasonable choice. In contrast, the general class of f-divergences [7] of probability distributions appears well-suited. For reasons

that will become clear presently, we choose the so-called χ^2-divergence defined as $\chi^2(\mathbf{p},\mathbf{p}') = \frac{1}{2}\sum_j \frac{(p_j - p'_j)^2}{p'_j}$. Like for the KLD, $\chi^2(\mathbf{p},\mathbf{p}') \geq 0$ with equality iff $\mathbf{p} = \mathbf{p}'$. It can then easily be shown that the NG update (9) is the solution of the proximal point problem

$$\mathbf{p}^{k+1} = \arg\max_{\mathbf{p}} \left\{ \tilde{I}^k(\mathbf{p}) - \gamma_k \chi^2(\mathbf{p},\mathbf{p}^k) \right\},$$

obtained with $\mu_k = 1/\gamma_k$. Thus, accelerated BA and NG can both be viewed as proximal point methods using the same cost function $\tilde{I}^k(\mathbf{p})$ but different distance measures for the proximity penalty. Their asymptotic equivalence follows from the well-known fact that $\chi^2(\mathbf{p},\mathbf{p}') \approx D(\mathbf{p}\|\mathbf{p}')$ for $\mathbf{p}$ close to $\mathbf{p}'$.

C. Choice of Step-Size. A fundamental property of the BA algorithm is that the mutual information $I(\mathbf{p}^k) = \sum_{j=1}^M p_j^k D_j^k$ which represents the current capacity estimate is non-decreasing. For the accelerated BA algorithm, it can be shown that

$$I(\mathbf{p}^{k+1}) \geq I(\mathbf{p}^k) + \gamma_k D(\mathbf{p}^{k+1}\|\mathbf{p}^k) - D(\mathbf{q}^{k+1}\|\mathbf{q}^k).$$

A sufficient condition for $I(\mathbf{p}^k)$ to be non-decreasing thus is $\gamma_k D(\mathbf{p}\|\mathbf{p}^k) - D(\mathbf{q}^{k+1}\|\mathbf{q}^k) \geq 0$, i.e., we have to ensure that

$$\mu_k \leq \frac{D(\mathbf{p}^{k+1}\|\mathbf{p}^k)}{D(\mathbf{q}^{k+1}\|\mathbf{q}^k)} = \frac{D(\mathbf{p}^{k+1}\|\mathbf{p}^k)}{D(\mathbf{Q}\mathbf{p}^{k+1}\|\mathbf{Q}\mathbf{p}^k)}. \qquad (14)$$

Motivated by the similarity to the squared maximum matrix eigenvalue $\sup_{\mathbf{x}\neq\mathbf{y}} \frac{d_E^2(\mathbf{A}\mathbf{x},\mathbf{A}\mathbf{y})}{d_E^2(\mathbf{x},\mathbf{y})}$ with $d_E^2(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2$, we define the *maximum KLD-induced "eigenvalue"* of $\mathbf{Q}$ as

$$\lambda_{\mathrm{KL}}^2(\mathbf{Q}) \triangleq \sup_{\mathbf{p}\neq\mathbf{p}'} \frac{D(\mathbf{Q}\mathbf{p}\|\mathbf{Q}\mathbf{p}')}{D(\mathbf{p}\|\mathbf{p}')}.$$

Note that $0 \leq \lambda_{\mathrm{KL}}^2(\mathbf{Q}) \leq 1$ and further $\lambda_{\mathrm{KL}}^2\left(\frac{1}{N+1}\mathbf{1}_{N+1}\mathbf{1}_{M+1}^T\right) = 0$, $\lambda_{\mathrm{KL}}^2(\mathbf{I}) = 1$. Thus, small $\lambda_{\mathrm{KL}}^2(\mathbf{Q})$ means that the channel is noisy. Using this definition, a sufficient condition for $I(\mathbf{p}^k)$ to be non-decreasing is given by

$$\mu_k \leq \frac{1}{\lambda_{\mathrm{KL}}^2(\mathbf{Q})}. \qquad (15)$$

In fact, we observed in our simulations that when using a fixed step-size, maximum convergence speed was achieved with $\mu = 1/\lambda_{\mathrm{KL}}^2(\mathbf{Q})$. Since a reasonable estimate of $\lambda_{\mathrm{KL}}^2(\mathbf{Q})$ might be difficult to obtain, we suggest to use the adaptive step-size $\mu_k = \frac{D(\mathbf{Q}\mathbf{p}^k\|\mathbf{Q}\mathbf{p}^{k-1})}{D(\mathbf{p}^k\|\mathbf{p}^{k-1})}$ in practical implementations. While this choice sometimes violates (15), we observed excellent performance (in fact, superlinear convergence) in our numerical experiments (see Section VII). The step-size μ_k could also be chosen using line-search techniques along the e-geodesics discussed in Section C. However, our experiments indicated that the proposed adaptive step-size yields similar performance at much lower complexity.

We note that the above arguments can also be used to choose the step-size of the NG algorithm since in the vicinity of the optimum solution accelerated BA and NG behave identical.

VI. Convergence Analysis

In the foregoing discussion we saw that step-sizes > 1 in the accelerated BA and NG algorithms have the potential for increased convergence speed. In this section, we summarize more explicit results regarding the convergence of the accelerated BA algorithm (full details are presented in [14]). While explicit results are difficult to obtain for the NG algorithm, Section IV and our simulations suggests that it inherits the convergence properties of the latter. We note that parts of our results in this section are inspired by [3, 5].

A. Convergence Statement 1. Consider the accelerated BA algorithm with $I^k = \sum_j p_j^k D_j^k$ and $L^k = \frac{1}{\mu_k}\log\left(\sum_j p_j^k \exp(\mu_k D_j^k)\right)$. Assume that $\mu_{\inf} \triangleq \inf_k \mu_k > 0$ and that (15) is satisfied for all k. Then, it can be shown that

$$\lim_{k\to\infty} L^k = \lim_{k\to\infty} I^k = C.$$

Furthermore, convergence rate is at least proportional to $1/k$,

$$C - L^k < c\frac{D(\mathbf{p}^*\|\mathbf{p}^0)}{\mu_{\inf}\, k}.$$

This clearly reflects that the accelerated BA algorithm ($\mu_k > 1$) converges faster than ordinary BA ($\mu_k = 1$).

B. Convergence Statement 2. Let us now assume that $\gamma_k = \gamma = 1/\mu$ is fixed. The fixed points $\mathbf{p}^*$ of accelerated BA and NG are defined as

$$\mathbf{p}^* = \arg\max_{\mathbf{p}} \left\{ \sum_j p_j D(\mathbf{Q}_j, \mathbf{Q}\mathbf{p}^*) - \gamma\, d(\mathbf{p}\|\mathbf{p}^*) \right\},$$

where $d(\mathbf{p}\|\mathbf{p}^k) = D(\mathbf{p}\|\mathbf{p}^k)$ for accelerated BA and $d(\mathbf{p}\|\mathbf{p}^k) = \chi^2(\mathbf{p},\mathbf{p}^k)$ for NG. This relation can be shown to be equivalent to (1) and thus $\mathbf{p}^*$ achieves capacity.

C. Convergence Statement 3. We next consider the accelerated BA algorithm and assume that the optimum input distribution $\mathbf{p}^*$ is unique, $p_j^* > 0$, and (15) is satisfied for all k.

For fixed step-size $\mu = 1/\gamma = 1/\gamma_k$, it then follows that the sequence $\mathbf{p}^k$ satisfies

$$\lim_{k\to\infty} \frac{\|\mathbf{p}^{k+1} - \mathbf{p}^*\|}{\|\mathbf{p}^k - \mathbf{p}^*\|} \leq \frac{c}{\mu},$$

i.e., the algorithm features (at least) linear convergence and convergence speed is increased by increasing μ.

Furthermore, it can be shown that there exists a step-size sequence $\gamma_k = 1/\mu_k$ conforming with (14) such that $\lim_{k\to\infty}\gamma_k = 0$ and

$$\lim_{k\to\infty} \frac{\|\mathbf{p}^{k+1} - \mathbf{p}^*\|}{\|\mathbf{p}^k - \mathbf{p}^*\|} = 0.$$

Thus, the accelerated BA algorithm with properly chosen step-size has the potential for superlinear convergence[1] (in fact we observed superlinear convergence with the adaptive step-size $\mu_k = \frac{D(\mathbf{Q}\mathbf{p}^k\|\mathbf{Q}\mathbf{p}^{k-1})}{D(\mathbf{p}^k\|\mathbf{p}^{k-1})}$).

[1] Newton-type methods could also be applied to achieve superlinear convergence. However, they require a matrix inversion and thus are computationally much more complex.

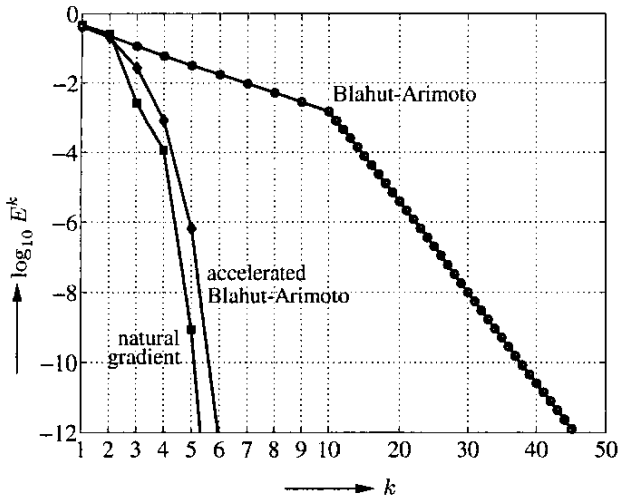

Figure 1: *Convergence of BA algorithm and accelerated BA and NG algorithms (note the nonuniform abscissa scaling).*

VII. Numerical Example

For purposes of illustration of our results, consider the channel $\mathbf{Q}^T = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.7 \end{pmatrix}$ from Fig. 4.5.1 in [11]. Let us ignore the fact that this channel can be recognized as being symmetric and that the optimum input distribution is thus uniform. To compute the capacity of this channel, we ran the BA algorithm, the accelerated BA algorithm, and the natural gradient algorithm with the same (randomly picked) initial guess $\mathbf{p}^0$. The step-size in our algorithms was chosen in an adaptive fashion as $\mu_k = D(\mathbf{Q}\mathbf{p}^k \| \mathbf{Q}\mathbf{p}^{k-1})/D(\mathbf{p}^k \| \mathbf{p}^{k-1})$ for $k > 1$ and $\mu_1 = 1$. As performance (and stopping criterion) we used $E^k = \max_j D_j^k - I^k$ since (2) implies $C - I^k \leq E^k$. The convergence results for a desired accuracy of 12 decimals are shown in Fig. 1. (all algorithms delivered $C = 0.365148445440$ bit). It is seen that the convergence of the NG and accelerated BA algorithms (6 iterations for the desired accuracy) is significantly faster than that of the BA algorithm (46 iterations). After five iterations, BA yields only one correct decimal while accelerated BA and NG achieve already 6 and 9 correct decimals, respectively. Fig. 1 also clearly verifies that the accelerated BA and NG algorithm with adaptive step-size feature superlinear convergence.

VIII. Conclusions

We have proposed improvements on the Blahut-Arimoto (BA) algorithm for computing the capacity of discrete memoryless channels (DMC). An accelerated BA algorithm and a natural gradient (NG) algorithm have been introduced that have the potential for significantly faster (in fact, often superlinear) convergence as compared to the conventional BA algorithm. Recasting the capacity computation problem as an equidivergence game, intuitive interpretations of these algorithms have been given via information geometric arguments. We also provided a unifying framework for the (accelerated) BA and NG algorithms in terms of proximal point methods. This allows for some statements regarding the convergence of our algorithms.

While our presentation focused on DMCs, our results carry over to the cases of multi-access channels [16], quantum channels [15], ISI channels [13, 17], and channels with side information [10]. In all of these cases, the computational savings achieved using our technique will be even more pronounced.

Furthermore, we conjecture that our approach can be applied to the computation of rate-distortion curves and to portfolio optimization, both of which represent problems closely related to capacity computation [4, 9].

References

[1] S. Amari and S. C. Douglas. Why natural gradient? In *Proc. IEEE ICASSP-98*, pages 1213–1216, Seattle, WA, May 1998.

[2] S. Amari and H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society and Oxford University Press, New York, 2000.

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

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

[5] S. Chretien and A. O. Hero. Kullback proximal algorithms for maximum-likelihood estimation. *IEEE Trans. Inf. Theory*, 46(5):1800–1810, Aug. 2000.

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

[7] I. Csiszár. Information type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[8] I. Csiszár and F. Matúš. Information projections revisited. *IEEE Trans. Inf. Theory*, 49(6):1474–1490, June 2003.

[9] I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions, Supplement Issue No. 1*, pages 205–237, 1984.

[10] F. Dupuis, W. Yu, and F. M. J. Willems. Blahut-Arimoto algorithms for computing channel capacity and rate-distortion with side information. In *Proc. IEEE ISIT 2004*, Chicago, IL, June/July 2004.

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

[12] D. P. Helmbold, R. E. Schapire, Y. Singer, and M. K. Warmuth. A comparison of new and old algorithms for a mixture estimation problem. *Machine Learning*, 27(1):97–119, 1997.

[13] A. Kavcic. On the capacity of Markov sources over noisy channels. In *Proc. IEEE GLOBECOM-2001*, pages 2997–3001, San Antonio, TX, Nov. 2001.

[14] G. Matz and P. Duhamel. Accelerating the Blahut-Arimoto algorithm via information geometry. *IEEE Trans. Inf. Theory*, to be submitted.

[15] H. Nagaoka. Algorithms of Arimoto-Blahut type for computing quantum channel capacity. In *Proc. IEEE ISIT 1998*, page 354, Cambridge, MA, Aug. 1998.

[16] M. Rezaeian and A. Grant. A generalization of the Arimoto-Blahut algorithm. In *Proc. IEEE ISIT 2004*, Chicago, IL, June/July 2004.

[17] P. O. Vontobel. A generalized Blahut-Arimoto algorithm. In *Proc. IEEE ISIT 2003*, page 53, Yokohama, Japan, June/July 2003.