Information Geometric Mechanisms Behind Maximising Channel Capacity

Kieran Morris

Contents

Abstract

Information Geometry provides a language to interpret information theoretic quantities and results in the language of geodesics, projections and metrics inherited from differential geometry. One problem which has been aided by this is computing channel capacity. In this paper we review the various information geometric interpretations of the Blahut-Arimoto algorithm, drawing attention to the shared structures present in each. We then go onto visualise the behaviour of the algorithm under multiple conditions, demonstrating more information geometric connections.

1 Channel Capacity

Consider an random (input) variable X on the a discrete alphabet $\mathcal{X} = \{1, ..., n\}$ with pdf $p(x)$, being fed through a noisy channel, giving us another random (output) variable Y on $\mathcal{Y} = \{1, ..., m\}$ with pdf $q(y)$. We can write

$$q(y) = \sum_{x \in \mathcal{X}} r(y \mid x) p(x) \tag{1}$$

where the conditional probability density function $r(y \mid x)$ is determined by the channel. Equivalently we can formulate the variables in terms of linear algebra which is more convenient computationally, if we define:

$$\begin{aligned} x_i &= \mathbb{P}(X = i) = p(i), \\ y_j &= \mathbb{P}(Y = j) = q(j), \\ R_{ij} &= \mathbb{P}(Y = j \mid X = i) = r(j \mid i), \end{aligned}$$

then we can express $y_j = \sum_k P_{kj} x_k$ which is equivalent to 7. We call $R = R_{ij}$ the *channel matrix.* The *capacity* of a channel $C = C(P)$ is defined by:

$$C = \max_X I(X : Y)$$

where I is the mutual information between X and Y - see [1, Page 184]. The mutual information can be defined as:

$$\begin{aligned} I(X, Y) &\coloneqq D(p_{XY} \mid\mid p_X p_Y) \\ &= D(p(x) r(y \mid x) \mid\mid q(y)), \end{aligned}$$

where D is the KL-divergence. We will sometimes write $I(X; P) = I(p(x); r)$ since Y is completely determined by r and X, or even omit the dependence on r, writing $I(X) = I(p)$ if it is obvious which r we are using throughout. The capacity of a channel is extremely important in information theory as it provides a lower bound for the rate at which meta data must be tacked onto transmissions to guarantee (asymptotic) lossless transmission [1, Page 198].

Certain analytic solutions are already known, for example for a binary symmetric channel, i.e where the matrix is of the form

$$R = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

then the *only* capacity achieving distribution is the uniform distribution, and the capacity is of the form

$$C = 1 - H(p).$$

Unfortunately we cannot always rely on analytic solutions, and must resort to numerical optimisation procedures. For these we have a constrained optimisation problem, since we must have $\sum_k x_k = 1$. It is known that the solutions for capacity achieving distributions forms a sub-simplex, [2], meaning we can take linear combinations of any solutions so long as they remain inside the original simplex. In fact we have an explicit characterisation of the solutions:

Given a single capacity achieving distribution p^*, the family of capacity achieving distributions is $P_X^* = \{p^* + \ker R\}$. In particular if $\ker R = \{0\}$ we have a unique solution.

1.1 The Blahut Arimoto Algorithm

One method of finding capacity is the Blahut-Arimoto algorithm [3], which formulates the single maximisation problem as two smaller maximisation problems, both of which have analytic solutions. One additional benefit of the Blahut-Arimoto algorithm is that it naturally stays restricted to the $(n-1)$ simplex.

We consider the following function:

$$I(X;W) = \sum_{i,j} R_{ij} x_i \ln \frac{W_{ji}}{x_i}.$$

where W_{ji} is a stochastic matrix. The following is true:

$$I(X;W) \leq I(X)$$

with equality if and only if

$$W_{ji} = \frac{R_{ij} x_i}{y_j} = \frac{R_{ij} x_i}{\sum_k R_{ik} x_k},$$

i.e the posterior probability $w(x \mid y) = r(y \mid x)p(x)/q(y)$ [3]. In terms of maximisation, we have $\max_W I(X;W) = I(X)$ for any distribution X.

Alternatively suppose we are given W_{ji}, and wish to find $\max_X I(X;W)$, then it can be shown via Lagrange multipliers that:

$$\text{argmax}_X\, I(X;W) = \frac{\exp[\sum_k R_{ki} \ln W_{ik}]}{\sum_k \exp[\sum_i R_{ki} \ln W_{ik}]}.$$

Giving us analytic solutions for both maximistions. It can also be shown that the process of alternately maximising over X and W is monotonically increasing and hence converges to a maximum [3]. See Algorithm 1 for pseudocode.

Algorithm 1: Blahut-Arimoto Algorithm

```
Input: Initial distribution $x^0$
Output: Capacity
$x \leftarrow x^0$ ;
while not converged do
    $W_{ij} \leftarrow R_{ij} x_j / \sum_k R_{kj} x_k$ ;
    $a_i \leftarrow \exp[\sum_k R_{ki} \ln W_{ik}]$ ;
    $x_i \leftarrow a_i / \sum_k a_k$;
return $I(x, W)$;
```

Thanks to this structure the algorithm is remarkably efficient at finding solutions to capacity. There have been multiple formulations of this algorithm, including improvements on convergence rate [4, 5] - which we will explore later on - and recent adaptations for quantum channels [6, 7, 8] relying heavily on the foundational algorithm. One final note is that the two maximisation steps can be combined analytically to give the following iterative formula for $p^{(t+1)}(x)$:

$$p^{(t+1)}(x) = \frac{p^{(t)}(x) \exp(D(r(y \mid x) \mid\mid q^{(t)}(y)}{\sum_x \exp(D(r(y \mid x) \mid\mid q^{(t)}(y))}. \quad (2)$$

We could of course write the linear algebra form, but this form will be particularly useful in this paper.

2 Preliminaries - Information Geometry

2.1 Objects and Structures

Information Geometry interprets an n-dimensional family of probability distributions $\mathcal{P}$ as a differential manifold. Allowing for the imposing of geometric structure on $\mathcal{P}$, this includes:

1. The metric tensor g: at each point p, we can compute the inner products on tangent vectors v, w based at any point $p \in \mathcal{P}$.
2. The dual coordinate frames: these are two global coordinate frames, i.e parametrisations of the family, denoted θ and η, these bases are linked by a Legendre transformation.

3. A divergence operator D: which is a positive definite measure of deviation between points, along with a dual divergence D^* which corresponds to flipping the arguments, i.e $D^*[p \mid\mid q] = D[q \mid\mid p]$.

4. The dual connections (or the dual Christoffel Symbols $\Gamma_{ij}^k(\theta)$ and $\Gamma_{ij}^{k*}(\eta)$): which describe how to translate vectors across a manifold via a differential equation.

5. Dual geodesics (or autoparallel curves): which interpolate between two distributions, we have one form of geodesic for each basis.

In practice, these abstract concepts boil down to familiar probabilistic concepts:

1. The Metric Tensor is the negative of the *Fisher Information Matrix*, i.e we have that $g(\theta) - \mathbb{E}[\partial_i l(x;\theta) \cdot \partial_j l(x;\theta)]$ where $l(x;\theta) = \log p(x;\theta)$ is the log-likelihood [9].

2. The first coordinate frame is typically already known, as it is the parameter vector θ which determines the probability distribution at each point. The second frame η can derived with a convex function ψ, by evaluating $\eta = \nabla\psi(\theta)$.

3. The divergence operator is the KL-divergence, and the dual connections are the third order derivatives as described.

4. The connections are again defined by the third order derivatives of D_{KL}.

5. The formulation of dual geodesics are described below.

Consider two points $p, q \in \mathcal{P}$ in both coordinate frames:

$$\begin{aligned} p &= (\theta_p) = (\eta_p) \\ q &= (\theta_q) = (\eta_q), \end{aligned}$$

then we can define the θ - geodesic, γ_θ, between them as the solution to the following differential equation:

$$\ddot{\gamma}^k(t) + \Gamma_{ij}^k \dot{\gamma}^i(t)\dot{\gamma}^j(t) = 0 \tag{3}$$

where Γ_{ij}^k are the Christoffel symbols uniquely defining the connection. Similarly we can define γ_η by replacing Γ_{ij}^k for Γ_{ij}^{k*}. In the case where the connection is flat, i.e $\Gamma_{ij}^k = 0$ for all i, j, k then the θ-geodesic between p and q is defined linearly by:

$$\gamma_\theta(t) = (1-t)\theta_p - t\theta_q.$$

Additionally, the Fundamental Theorem of Information Geometry [10, Page 13], states that $\Gamma_{ij}^k = 0$ if and only if $\Gamma_{ij}^{k*} = 0$, hence an η-geodesic is similarly defined by:

$$\gamma_\eta(t) = (1-t)\eta_p - t\eta_q.$$

Note that while in their own coordinate frame, the geodesics may appear linear, however in the θ frame, γ_η may appear far from linear, and vice versa. We say a structure like this is *dually flat* - in that they are flat in two different frames. The next section focuses on a very large family of dually flat manifolds - the exponential families.

2.2 Exponential Families

The class of exponential families are a large family of distributions which have the following *canonical* form:

$$p(x;\theta) = \exp(t(x)\cdot\theta + k(x) - \psi(\theta)), \quad \theta \in \Theta,$$

where we call the $\theta = (\theta^i)$ the *natural paramters*, $t(x) = (t_i(x))$ is a vector function of x and $\psi(\theta)$ is the (convex) normalising factor. The exponential families lend themselves particularly well to the world of information geometry, lets compute our important objects for exponential families.

1. Computing the metric tensor for the exponential family we get: $g(\theta) = \nabla^2\psi(\theta)$.

2. For an exponential family, the dual frame can be derived via the *expectation parameter*, which is a vector defined by $\eta_i(\theta) = \mathbb{E}_p[t_i]$. However a more useful form can be shown to be: $\eta(\theta) = \nabla_\theta\psi(\theta)$ satisfying our original definition. Additionally we get that $\theta = \nabla\varphi(\eta)$ where φ is the negative Shannon entropy of $p(x;\theta)$. Giving us dual structure between (θ, ψ) and (η, φ) with the condition $\varphi(\eta) + \psi(\theta) = \theta\cdot\eta$ [11, Page 139].

3. We can see that evaluating the KL divergence between two elements of an exponential family θ, θ', gives

$$D_{KL}[\theta \mid\mid \theta'] = \psi(\theta) - \eta'\cdot\theta + \varphi(\eta')$$

where φ is again the negative entropy function [11].

4. Since D_{KL} can be expressed as a quadratic combination of functions of θ and η, the third order derivatives are all zero [10], and we have dually flat geometry.

5. This means that geodesics (autoparallel curves) for exponential families are linear as previously discussed.

Example 1 (Normal Family)
The family of normal distributions is an exponential family with the following assignment:

$$k(x) = 0, \quad t_1(x) = x, \quad t_2(x) = x^2,$$
$$\theta_1 = \frac{\mu}{\sigma^2}, \quad \theta_2 = \frac{-1}{2\sigma^2},$$
$$\psi(\theta) = \frac{-(\theta_1)^2}{2\theta_2} + \frac{1}{2}\ln\left(\frac{-\pi}{\theta_2}\right).$$

Then the metric is:

$$g_{11} = -\frac{1}{\theta_2}, \quad g_{22} = \frac{\theta_2 - (\theta_1)^2}{2(\theta_2)^3}$$
$$g_{12} = g_{21} = \frac{\theta_1}{4(\theta_2)^2}.$$

The dual basis $\eta = \nabla_\theta \psi(\theta)$ is:

$$\eta = \left(\frac{-\theta_1}{2\theta_2}, \frac{(\theta_1)^2 - 2\theta_2}{4(\theta_2)^2}\right)$$
$$= (\mu, \mu^2 + \sigma^2)$$

From this point we can define both types of geodesic, consisting of linear interpolation in each frame.

The following example will be particularly useful for our purposes, as it is the space of probability distributions that Arimoto works over.

Example 2 (Discrete Distributions)
Suppose we have a discrete distribution over $\mathcal{X} = \{0, 1, ..., n\}$ in the following form:

$$p(x) = \sum_{i=0}^{n} p_i \delta_i(x),$$

so that $p(i) = p_i$. We see that this is an exponential family via:

$$p(x; \theta) = \exp\left(\sum_{i=1}^{n} \theta^i \delta_i(x) - \psi(\theta)\right)$$

where $\theta^i = \ln \frac{p_i}{p_0}$ and $\psi(\theta) = -\ln(1 + \sum_{i=1}^{n} \exp(\theta^i))$. It can also be shown that the dual coordinate frame is $\eta = (p_1, ..., p_n)$, so we have that

$$\theta^i = \ln \frac{\eta_i}{1 - \sum \eta_i} \tag{4}$$

Additionally we have the notion of a *mixture family*, which are of the form:

$$p(x, \eta) = \sum_{i=0}^{n} e_i q_i(x)$$

for $\{q_i\}$ linearly independent and $\sum_{i=0}^{n} e_i = 1$. Clearly the discrete distributions are a mixture family using the dual basis, setting $\eta_i = e_i$ and $q_i = \delta_i$. Meaning we can interchangeably consider a discrete distribution as either an exponential family, or a mixture family - which are dually linked together.

This leads to more concrete names for our coordinate frames, we refer to θ as the exponential frame, and its geodesic as an e-geodesic, and similarly we refer to η as the mixture frame, and its geodesic as an m-geodesic.

2.3 Pythagorean and Projection Theorems

For this section we explore some theorems which parallel results in flat Euclidean geometry.

Theorem 1 (The Pythagorean Theorem)
Given three points $p, q, r \in \mathcal{P}$, let γ_1 be the e-geodesic connecting p and q and γ_2 be the m-geodesic connecting q and r, then if γ_1 and γ_2 are orthogonal (i.e $g_q(\dot{\gamma_1}, \dot{\gamma_2}) = 0$) then we have that

$$D(p \mid\mid r) = D(p \mid\mid q) + D(q \mid\mid r).$$

We equivalently have a dual Pythagorean theorem for the dual divergence, which is defined by switching the orders of the argument in D, i.e $D^*(p \mid\mid q) = D(q \mid\mid p)$, and the role of the e/m-geodesics are flipped, see [11, Theorem 1.2,1.3].

Next we consider two types of submanifold, m-autoparallel and e-autoparallel submanifolds, which play the same role as affine subspaces in Euclidean geometry, they are defined abstractly, see [12, Theorem 1.1], but the following sufficient condition will suffice for our purposes.

Theorem 2 (Autoparallel Submanifolds)
For a submanifold $M \subset \mathcal{P}$, if M is convex over θ, then M is e-autoparallel, and if M is convex over η, then M is m-autoparallel, [13, Theorem 2.6].

Autoparallel submanifolds have the property that projections are unique.

Theorem 3 (Projection Theorem)
Given a point $p \in \mathcal{P}$ and an m-autoparallel submanifold M, then a necessary and sufficient condition for a point $q \in M$ to satisfy $\min_{r \in M} D(p \mid\mid r)$ is for the e-geodesic between p and q to be orthogonal (with respect to the metric) to M.

See [12, Theorem 3.9]. This is one of the more pragmatic theorems in information geometry, and is one that we will see very often throughout this paper.

3 The Geometry of Blahut-Arimoto

To say *the* geometry is a little misleading, as there are many ways to interpret its geometry. Many of which stem from the equivalence of minimising KL-divergence and projection via an e or m-geodesic, the differences lie in how the KL-divergence is brought into the equation.

3.1 Hints of a Projection

An early interpretation comes from an example in [14], a paper on a family of alternating minimization algorithms, also explored in [11]. Csiszar expresses two families:

$$\mathcal{P} = \{p(x) \cdot r(y \mid x) : p \in S_n\},$$
$$\mathcal{Q} = \{w(x \mid y) r(y \mid x) : w \text{ is a measure on } x \}$$

Then the channel capacity is

$$\begin{aligned} C(P) &= -D(\mathcal{P} \mid\mid \mathcal{Q}) \\ &:= - \min_{p \in \mathcal{P}, q \in \mathcal{Q}} D(p \mid\mid q) \end{aligned}$$

and then by utilising a looser definition of a projection, we can see we are performing alternating projections between $\mathcal{P}$ and $\mathcal{Q}$. Unfortunately $\mathcal{Q}$ consists of measures, not necessarily probability distributions, and the projections are not necessarily unique. More recent interpretations provide a more concrete description which align more with modern formulations.

One benefit of this interpretation is that it is not limited to channel capacity problems, in [14] itself we see other examples of capacity per unit cost, rate distortion and log-portfolio maximisation. Later works also showed that it can be applied to EM algorithms [1, Page 346].

3.2 Alternating Projections

3.2.1 Geometry of Capacity

The strongest example which utilises the structure of Information Geometry comes from [13], in a similar vein to [14] they similarly define two families:

$$\mathcal{M} = \{p(x) \cdot r(y \mid x) : p \in S_n\}, \tag{5}$$
$$\mathcal{E} = \{p(x) \cdot q(y) : p \in S_n, q \in S_m\} \tag{6}$$

where S_k denotes the set of discrete distributions on k variables. Note that $\mathcal{M}, \mathcal{E} \subset S_{nm}$. In [13] we have a more explicit statement of our condition for submanifolds of exponential families to be autoparallel:

If for any $p_1, p_2 \in M$ and $t \in (0,1)$, we have that $p_3 := tp_1 + (1-t)p_2 \in M$, then M is (m)-autoparallel. Similarly if for any $p_1, p_2 \in\ E$ and we define p_3 by: $\log p_3 := t \log p_1 + (1-t) \log p_2 + A \in E$ (where A is a normalising constant) then if $p_3 \in \mathcal{E}$ for all $t \in (0,1)$ then $\mathcal{E}$ is (e)-autoparallel.

From this we see that $\mathcal{M}$ is (m)-autoparallel since

$$tp_1 \cdot r + (1-t)p_2 \cdot r = [tp_1 + (1-t)p_2] \cdot r \in \mathcal{M}$$

and $\mathcal{E}$ is (e)-autoparallel since

$$\begin{aligned} t\log(p_1 \cdot q_1) + (1-t)\log(p_2 \cdot q_2) &= \\ t \log p_1 + (1-t)\log p_2 + t \log q_1 + (1-t) \log q_2 &= \\ \log p_3 + \log q_3 &= \\ \log(p_3 \cdot q_3) & \end{aligned}$$

for $p_3, q_3 \in \mathcal{E}$ - note in both of these we are using that S_k is closed under convex combinations of θ and η. Thus we have unique projections from any $p(x,y) \in S_{nm}$ onto $\mathcal{M}$ and $\mathcal{E}$. In particular we have that the m-projection of $p(x)r(y \mid x)$ from to $\mathcal{E}$ is $p(x) \times q(y)$. Then we can see that the capacity:

$$C = \max_{p(x,y) \in \mathcal{M}} D(p(x,y) \mid\mid p(x) \times q(y))$$

is the maximum divergence between $p(x,y) \in \mathcal{M}$ and its projected point in $\mathcal{E}$. To expand this interpretation to Blahut-Arimoto, we must discuss the *backwards em-algorithm.*

3.2.2 The Backwards em Algorithm

The standard *em*-algorithm is the process alternatively performing e and m-projections between two submanifolds of probability distributions [15]. These have many applications, including estimation from data in neural networks [11, 16]. We see the pseudo-code in Algorithm 2.

```
Algorithm 2: em-Algorithm
Input: Initial $p^0$, submanifolds $M$,$E$
$p \leftarrow p^0$;
while not convergeed do
    $u \leftarrow m$-project to $E$;
    $p \leftarrow e$-project to $M$;
return $p, u$;
```

In fact under certain conditions, the *em*-algorithm aligns exactly with the EM-algorithm [15], [11, Pages 183-184].

In [13], Shoji constructs a reverse *em*-algorithm, which instead of projecting from a point $p \in \mathcal{M}$ to $q \in \mathcal{E}$ via an m-projection we find the point p which projects to q, this is called a *backwards m*-projection. Similarly we can define a backwards *e*-projection. In Algorithm 3 we see pseudo-code for this.

Algorithm 3: Backwards em-Algorithm

```
Input: Initial distribution $u^0 \in M$
$u \leftarrow u^0$ ;
while not converged do
    $v(x,y) \leftarrow p(x) \times q(y)$ s.t $p(x) \times q(y)$ $(e)$-projects to $u$ ;
    $u(x,y) \leftarrow p(x)r(y \mid x)$ s.t $p(x)r(y \mid x)$ $(m)$-projects to $v$ ;
return $u(x,y)$;
```

Note however, that this pseudo-code does not give us a way to find these reverse projections, or if they are unique, or even if they exist!

For some $p^t \in \mathcal{M}$, we attempt the backwards *e*-projection, we have an exponential family of possible solutions which Shoji calls $E^{(t)}$, it can be shown that this is an exponential family and is hence *e*-autoparallel [12, Theorem 5.3]. We don't yet have a way to choose which element is ideal, however we do know that it must be in the image of the *m*-projection - otherwise how could we reverse it!

The elements of this intersection are described by an intractable - and admittedly messy - equation, see [13, Eq. 16]. Shoji shows that updating p with solutions to this equation monotonically increases the mutual information and converge to capacity.

3.2.3 Approximation of the backwards *em*-algorithm

While it is possible that the exact solutions may be approximated via numerical methods - which seems to as of yet not been explored - Shoji goes on to make the following approximation:

$$r(y \mid x) \approx q^*(y),$$

where $q^*(y)$ is the distribution which achieves channel capacity. Remarkably, using the distribution which achieves capacity for the approximation does not mean we have to compute it. This gives us a much simpler equation to solve for the backwards *e*-projection [13, Eq. 17]. Shoji then performs an *m*-projection as an approximation of a backwards *m*-projection to give the following iterative formula for the approximate backwards *em*-algorithm:

$$p^{(t+1)} \propto p^{(t)} \exp[D(r(y \mid x) \mid\mid q^{(t)}(y)]$$

which is exactly the Blahut-Arimoto algorithm! Meaning that each pair of maximisations over W and X corresponds to an approximate backwards (*em*) projection.

This interpretation contains all the hallmarks of information geometry, thanks to it's origin as an adaptation of an *em*-algorithm, making it the strongest interpretation in this paper. It also naturally extends from Csiszar's initial interpretation, leading us to speculate if similar extensions could be built from the other examples of alternating minimisations given in [14].

3.3 Singular Projection

Another interpretation, found in Naja, Alberge and Duhamel's [5] is to view each update (to p^t) as a minimisation procedure. They show that

$$p^{(t+1)} = \begin{cases} \min_p D(p(x) \mid\mid p^{(k)}) \\ \text{s.c} \quad I^{(k)}(p(x)) = \alpha \\ \text{s.c} \quad \sum_x p(x) = 1 \end{cases}$$

where $I^{(k)}(p(x))$ is shorthand for the current capacity estimate, has the following solution:

$$p^{(k+1)} = \frac{p^{(k)} \exp(\lambda_1 D_x^k)}{\sum_x p^{(k)} \exp(\lambda_1 D_x^k)} \tag{7}$$

where $D_x^k = D(r(y \mid x) \mid\mid q^k(y))$. Which corresponds exactly with Blahut-Arimoto when $\lambda_1 = 1$ - in fact, in [4, 5] authors demonstrate that for certain values of λ_1 we have considerably faster convergence. Looking back at the projection theorem, at each step, we are projecting onto the submanifold for which $I^{(k)}(p) = \alpha$ for all p. As authors state however, this doesn't tell us what choice of α to make, at each stage, we only know that we must choose increasing α each time. The geometry of this interpretation is a useful addendum, however computational result that modifying λ_1 leads to faster convergence - with minimal computational cost - is the main takaway here.

3.4 Natural Gradient

Contained in another Duhamel paper: [4] is an adapted natural gradient method which aims to achieve channel capacity, their formula is:

$$p^{(t+1)}(x) = p^t(x)[1 + \lambda_k(D_x^k - I^k(p)] \tag{8}$$

Authors show that a taylor expansion approximation of (7) gives (8) exactly, with step size $\lambda_k = \lambda_1$. As mentioned before, this leads to faster convergence in what authors call the *Accelerated Arimoto Algorithm* by varying the λ_k parameter. This is another example of some minor geometry leading to asymptotically faster convergence of Blahut-Arimoto, since the natral gradient method is an adaptation of Gradient Descent which uses the Fisher Information Matrix [11, Page 282].

3.5 Proximal Point

Again in [5], we have another interpretation as a proximal point algorithm. Recall that a proximal point method simply refers to an optimisation problem which has a penalty for deviating too far from the previous iteration. They use the following formula:

$$p^{(t+1)} = \text{argmax}_p[I(p) - \lambda_k(D(p \mid\mid p^{(k)}) \\ -D(q \mid\mid q^{(k)}))],$$

where $q = q(y)$ and $q^{(k)} = q^{(k)}(y)$ are the corresponding distributions on Y associated to $p = p(x)$ and $p^{(k)}(x)$, and show it is equivalent to the Blahut-Arimoto algorithm. Here we are maximising the mutual information, while using the KL-divergence between the new p and new q as a penalty term. The λ_k term plays the same role as before and can lead to faster convergence.

The limit of the theory for this interpretation is using D_{KL} as a measure of deviation which is not particularly new, but it's connection to the previous interpretations and their improvements is worth mentioning.

3.6 Smallest Enclosing Circle Algorithm

Besides Arimoto there are other algorithms which reconfigure the capacity maximisation problem as a double optimisation problem too. The following algorithm, found in [17], reformulates the capacity as a min-max problem, noticeably different from the max-max Arimoto algorithm and it's alternative formulations. More specifically the capacity can be formulated as

$$C = \min_X \max_{i \leq m} D(R^i \mid\mid X)$$

where R_i is the i'th row of R - see [18]. Then noticing this problem is remarkably similar to a min-max problem over the reals:

$$C = \min_{x \in \mathbb{R}^n} \max_{1 \leq i \leq m} d(r^i, x)$$

where r_i are arbitrary n-ary vectors. The solution to this problem is known to be the *smallest enclosing circle*. Authors construct an algorithm that utilises projections to solve the euclidean smallest enclosing circle problem, and then almost line by line reconstruct it under the information geometric formulation to solve the capacity problem.

4 Simulations

In the theme of a geometric view, we finish this study with analysis of the behaviour of the Arimoto algorithm, with a conjecture about high dimensional results. The python code for these simulations can be found at **github.com/NonDescriptMaths/Arimoto**. Including a fully functional implementation of the Arimoto algorithm in Python and C++; generators for random channel matrices and priors; a function to check whether the channel will have either a convex or unique solution; and functions for analysis of the convergence and distribution of solutions for the algorithm. Of these functions many depend on the **geomstats** package [19, 20, 21], which allows to information geometric computation like Fisher Geodesics, autoparallel curves, and divergences.

4.1 Unique Solutions

We begin by considering the case where $\ker R = \{0\}$, so that the solutions for the BA algorithm are equivalent to the capacity attaining distribution. Additionally we have no need to worry about which prior we sample from for our initial distributions, this will have an impact in the next section.

4.1.1 Convergence

Consider the symmetric channel defined by the following matrix:

$$R = \begin{bmatrix} 2/3 & 1/6 & 1/6 \\ 1/6 & 2/3 & 1/6 \\ 1/6 & 1/6 & 2/3 \end{bmatrix} \tag{9}$$

which we can show analytically has a capacity achieving distribution at $p = (1/3, 1/3, 1/3)$, Figure 1 demonstrates BA's convergence. The main observation is that that the BA algorithm does not follow the geodesic at all. If we were using natural gradient methods, we would expect to see the points follow the geodesic exactly. Figure 2 contains a few other examples with varying channel matrices.

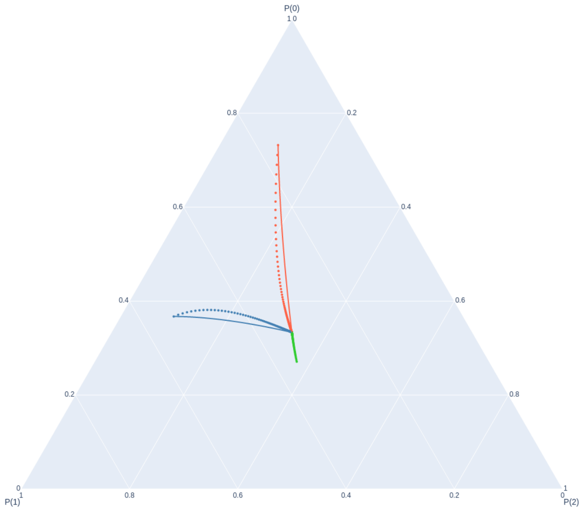


Figure 1: The convergence of BA to the uniform distribution with channel matrix (9) with three initial priors (dotted). Also included are the Fisher Information Geodesics (lined) for comparison.

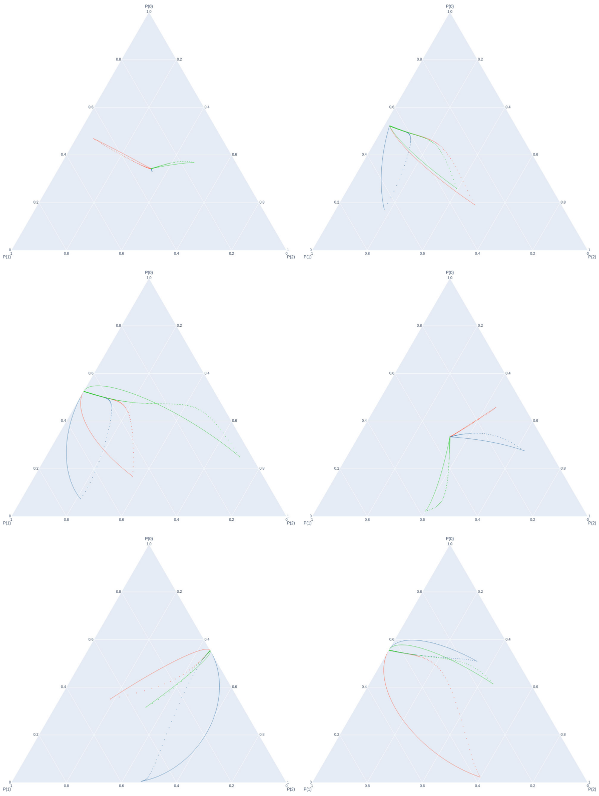


Figure 2: Blahut-Arimoto convergence for uniformly generated channel matrices with input dimension 3, and variable output dimension, alongside the geodesic from the Fisher Information Matrix.

4.1.2 Distribution

In the case for unique solutions, what is the distribution of solutions for a randomly selected channel? In Figure 3 we see the distribution for the 3 × 3 channel matrices. Considering we can easily visualize a 3 dimensional input alphabet, we varied the output alphabet and generated 1×10^6 uniformly generated channel matrices for each. The results can be seen in 4.

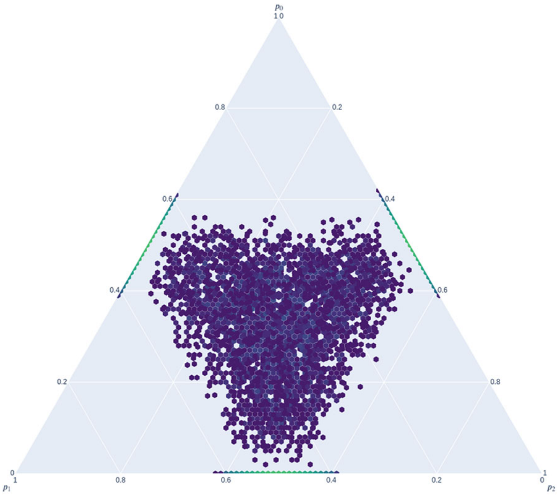


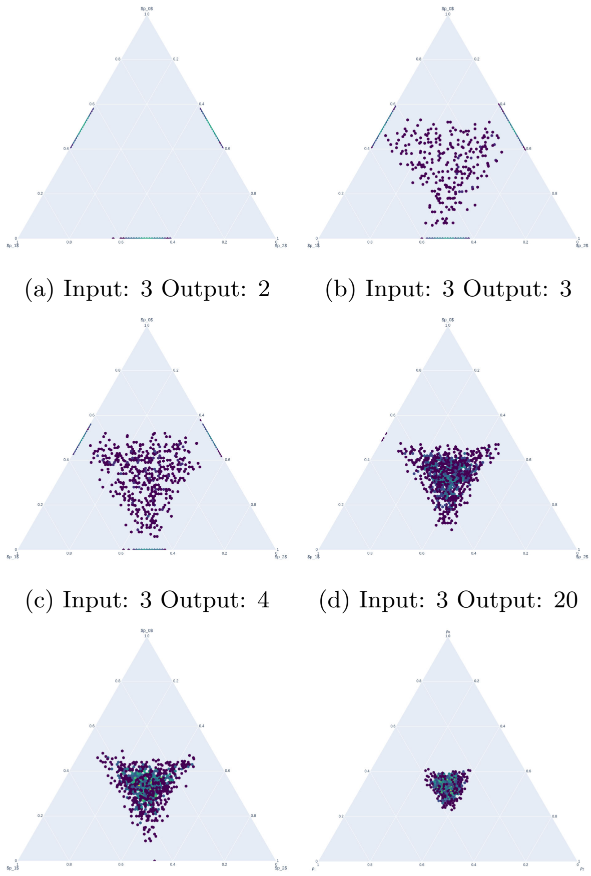
Figure 3: A heat-map for the convergence of Blahut-Arimoto for uniformly generated 3×3 channel matrices, each with a unique solution.

Some interesting behaviour - which was also present in Figure 2 - is that for 3 × 2, 3 × 3 and 3 × 4 we have a substantial amount of points on the boundary of the simplex, and in the 3 × 2 case it seems we only have convergence on the boundary, which may be a result that can be analytically proven in later work. Additionally, as the output dimension increases, we see that the capacity achieving solutions tend towards the uniform distribution - which again is worth investigating analytically.

4.2 Convex Solutions

We now consider the scenario where we do not have a unique solution. This means we can no longer make claims about the structure of capacity solutions, rather we can discuss the behaviour of BA itself - which debatably is more appropriate for this paper. Consider the channel defined by the following matrix:

$$R = \begin{bmatrix} 1/4 & 1/4 & 1/2 \\ 1/4 & 1/4 & 1/2 \\ 1/2 & 0 & 1/2 \end{bmatrix} \qquad (10)$$



(a) Input: 3 Output: 2

(b) Input: 3 Output: 3

(c) Input: 3 Output: 4

(d) Input: 3 Output: 20

(e) Input: 3 Output: 40

(f) Input: 3 Output: 160

Figure 4: A heat-map for the convergence of Blahut-Arimoto for uniformly generated $3 \times y$ channel matrices as y varies, each with a unique solution.

We can immediately see that $\ker(R)$ will be non-trivial since the first two rows are identical, in fact it can be shown that $\ker(R) = \text{span}\{(-1, 0, 1)\}$ and that P_X^* is equal to the convex hull of $\{(0.4, 0, 0.6), (0, 0.4, 0.6)\}$. We can also generate permutations of this channel by permuting the bottom row, this corresponds to permuting $(-1, 0, 1)$ and in turn rotating the set of solutions around the simplex, see Figure 6 for the distribution.

4.2.1 Convergence

Interestingly enough, when we plot the path of convergence for the convex set induced by 10, the BA algorithm happens to follow the geodesic exactly, see Figure 5. It in unclear as of yet whether this is a mathematical fluke or some underlying structure.

4.2.2 Distribution

Since the solutions are not unique, we choose to vary the method in which we sample the prior distributions, as these will clearly have an impact on the distribution of solutions. We begin with what we call the *niave* method, where we uniformly generate 3 numbers via the **random** package in python, then dividing by the sum of the numbers to make them sum to 1. It is clear that this will not be uniform and have bias towards a the uniform distribution [22].

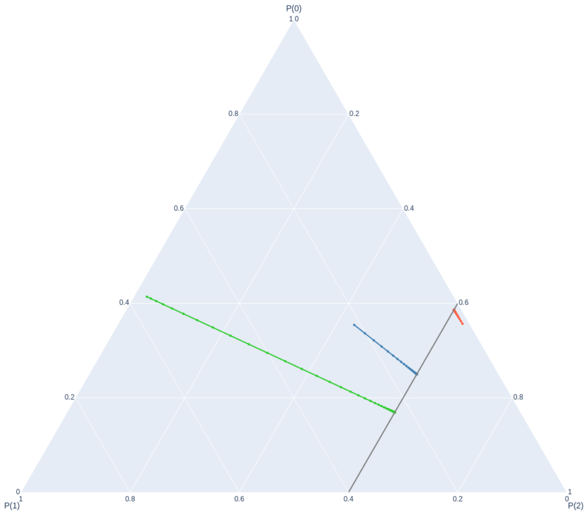


Figure 5: Convergence of BA to the convex set of solutions with the channel matrix from Eq 10 with three initial priors (dotted. Also included are the Fisher Information geodesic (lined) and the set of solutions (grey line).

The more standard second method is done by sampling from a (uniform) Dirichlet distribution on 3 variables.

The impact of the naive prior is seen in Figure 6 and the Dirichlet prior in 7. In these figures we can see that the dirichlet prior converges uniformly across the convex set of solutions, whereas the niave prior is more concentrated in the center of the set. This behaviour is consistent with our observations that the Blahut-Arimoto algorithm 'projects' from its initial position down to the convex set via a Fisher Geodesic - note the term projects here is not to be confused with our definition in section 2.

5 Conclusions

In this paper we have discussed the fundamental structures of information geometry and how it's techniques can be used in information theoretic problems, taking special interest in how the Blahut-Arimoto algorithm has multiple formulations in terms of projections. We also demonstrated that the behaviour of BA under multiple assumptions

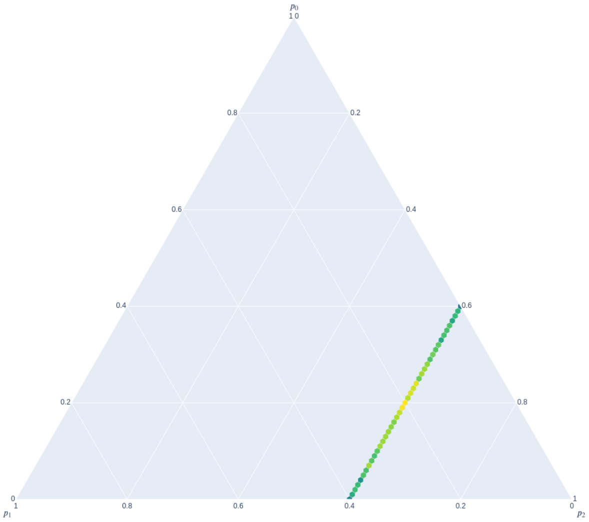


Figure 6: Distribution of BA solutions with a niave prior for channel capacity induced by Eq 10.

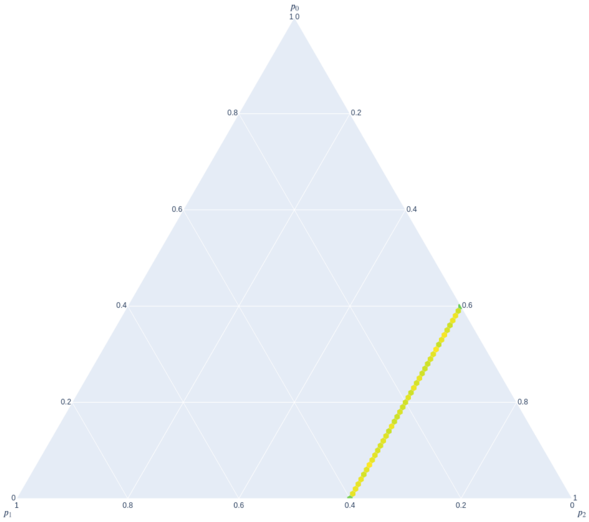


Figure 7: Distribution of BA solutions with a Dirichlet prior for channel capacity induced by Eq 10.

can vary greatly and possibly give rise to even more information geometric results. This paper did not directly present the cases of high dimension input alphabet which we are sure have similar intriguing properties.

5.1 Future Work

As stated in the simulations section, we have many computational results that require analytic investigation, which may bring insight into the behaviour of the algorithm - specifically the behaviour of high dimensional output alphabets on solutions, and the reason for BA 'projecting' onto sets of convex solutions.

As briefly mentioned, an attempt to find an exact or other approximate solution to the condition in [13] has not yet been explored - it may be worth investigating the accuracy/performance of numerical methods on the backwards *em*-algorithm. Additionally, considering it's structure is not limited to the families $\mathcal{M}$ and $\mathcal{E}$, generalisation to other autoparallel families seems possible - in particular the other examples of alternating projections in [14] such as capacity per unit cost and rate-distortion functions.

Considering there have been many publications in the field of quantum channels recently [8, 6, 7], it may be worth exploring the literature a little deeper and seeing its relation to information geometry - there are already publications in relation to quantum information geometry [23].

Considering our observations from the simulations, it's also worth investigating the KKT conditions for high dimensional y to see if there is a theoretical justification for the behaviour.

References

[1] T. M. Cover, *Elements of information theory.* John Wiley & Sons, 1999.

[2] Q. Ding, S. Jaggi, S. Vatedka, and Y. Zhang, “Empirical properties of good channel codes,” in *2020 IEEE International Symposium on Information Theory (ISIT).* IEEE, 2020, pp. 2337–2342.

[3] S. Arimoto, “An algorithm for computing the capacity of arbitrary discrete memoryless channels,” *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[4] G. Matz and P. Duhamel, “Information geometric formulation and interpretation of accelerated Blahut-Arimoto-type algorithms,” in

Information theory workshop. IEEE, 2004, pp. 66–70.

[5] Z. Naja, F. Alberge, and P. Duhamel, "Geometrical interpretation and improvements of the Blahut-Arimoto's algorithm," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 2009, pp. 2505–2508.

[6] N. Ramakrishnan, R. Iten, V. Scholz, and M. Berta, "Quantum Blahut-Arimoto algorithms," in *2020 IEEE International Symposium on Information Theory (ISIT).* IEEE, 2020, pp. 1909–1914.

[7] N. Ramakrishnan, R. Iten, V. B. Scholz, and M. Berta, "Computing quantum channel capacities," *IEEE Transactions on Information Theory*, vol. 67, no. 2, pp. 946–960, 2020.

[8] M. Hayashi and G. Liu, "Generalized quantum Arimoto-Blahut algorithm and its application to quantum information bottleneck," *arXiv preprint arXiv:2311.11188*, 2023.

[9] J. G. Dowty, "Chentsov's theorem for exponential families," *Information Geometry*, vol. 1, no. 1, pp. 117–135, 2018.

[10] F. Nielsen, "An elementary introduction to information geometry," *Entropy*, vol. 22, no. 10, p. 1100, 2020.

[11] S. ichi Amari, *Information Geometry and Its Applications.* Springer, 2015.

[12] S.-i. Amari and H. Nagaoka, *Methods of information geometry.* American Mathematical Soc., 2000, vol. 191.

[13] S. Toyota, "Geometry of arimoto algorithm," *Information Geometry*, vol. 3, no. 2, pp. 183–198, 2020.

[14] I. Csiszár, "Information geometry and alternating minimization procedures," *Statistics and Decisions, Dedewicz*, vol. 1, pp. 205–237, 1984.

[15] S.-i. Amari, "The em algorithm and information geometry in neural network learning," *Neural Computation*, vol. 7, no. 1, pp. 13–18, 1995.

[16] ——, "Information geometry of the em and em algorithms for neural networks," *Neural networks*, vol. 8, no. 9, pp. 1379–1408, 1995.

[17] K. Nakagawa, K. Watabe, and T. Sabu, "On the search algorithm for the output distribution that achieves the channel capacity," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 1043–1062, 2016.

[18] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems.* Cambridge University Press, 2011.

[19] N. Miolane, N. Guigui, A. Le Brigant, J. Mathe, B. Hou, Y. Thanwerdas, S. Heyder, O. Peltre, N. Koep, H. Zaatiti *et al.*, "Geomstats: a python package for riemannian geometry in machine learning," *Journal of Machine Learning Research*, vol. 21, no. 223, pp. 1–9, 2020.

[20] N. Miolane, N. Guigui, H. Zaatiti, C. Shewmake, H. Hajri, D. Brooks, A. Le Brigant, J. Mathe, B. Hou, Y. Thanwerdas *et al.*, "Introduction to geometric learning in python with geomstats," in *SciPy 2020-19th Python in Science Conference*, 2020, pp. 48–57.

[21] A. Le Brigant, J. Deschamps, A. Collas, and N. Miolane, "Parametric information geometry with the package geomstats," *ACM Transactions on Mathematical Software*, vol. 49, no. 4, pp. 1–26, 2023.

[22] N. A. Smith and R. W. Tromble, "Sampling uniformly from the unit simplex," *Johns Hopkins University, Tech. Rep*, vol. 29, 2004.

[23] J. Lambert and E. Sørensen, "From classical to quantum information geometry: a guide for physicists," *New Journal of Physics*, vol. 25, no. 8, p. 081201, 2023.