A Completed Information Projection Interpretation of Expectation Propagation

John MacLaren Walsh
Department of Electrical and Computer Engineering
Drexel University
Philadelphia, PA 19104
jwalsh@ece.drexel.edu

Abstract

Expectation propagation (EP), a family of methods for iterative approximate statistical inference closely related to belief propagation, is linked to a hybrid between Dykstra's algorithm with cyclic Bregman projections and the method of alternating Bregman projections from convex analysis via the use of the information geometry of exponential families. Doing so justifies extrinsic information extraction within the context of projections on convex sets, without the need for an iteration varying convex set which was required in previous information geometric descriptions of EP. It is suggested that new convergence results for EP might be developed through this connection by adapting the convergence proofs for alternating projections and Dykstra's algorithm with cyclic Bregman projections.

1 Introduction

Expectation propagation (EP) under reciprocity and sufficiency is a family of distributed iterative methods for statistical inference which includes the forward backward algorithm, the Kalman filter, Gallager's soft decoding algorithm for low density parity check (LDPC) codes, and the turbo decoder and as special cases. When EP corresponds to belief propagation (BP) and the factor graph which BP may be interpreted as a message passing algorithm on has no loops, which includes the forward backward algorithm and Kalman filter special cases, it is easy to show that EP correctly provides marginal a posteriori distributions after a finite number of iterations. When there are loops in the graph, which include the recent advanced applications to turbo and LDPC codes, however, the performance and convergence of the algorithms remain less understood. Furthermore, when the approximating family of distributions in EP does not match the family that the marginal distribution of the density being approximated, (differentiating EP from BP) the performance and convergence of EP is not completely understood.

Several authors have noted that belief/EP form a family of methods whose stationary points correspond to critical points of an approximate free energy. Specialists in information geometry subsequently noted that several steps of the algorithm admit direct interpretations as Kullback-Leibler divergence minimizing information projections. A complete interpretation of the algorithm family as a traditional alternating projection algorithm between time invariant convex sets, however, has remained incomplete due to an interfering step in the iteration, intrinsic information extraction, which may not be interpreted as an information projection on a time-invariant set. Here we provide a possible completion of the interpretation of EP as an iterative projection algorithm on time-invariant sets by linking EP with a hybrid between Dykstra's algorithm with cyclic Bregman projections and alternating Bregman projections from the convex analysis, optimization, and programming literature.

We are not the first to note the relevance of projection algorithms to specific instances of EP. Indeed, as phrased at its inception, EP was formulated with information (Kullback Leiber distance) minimizing projections, but the exposition used (depicted in (4) below) requires the formation of intermediate densities v_a whose formation, involving the removal from the approximate density of the factor g_a henceforth (non-traditionally) called intrinsic information extraction, is not well described as a projection on a time invariant set of densities. Within the domains of special cases of EP, very early on [1] noted the relevance of information projections to turbo decoding. Later, [2, 3] and [4] considered belief propagation and turbo decoding to be projection algorithms on information sets. These later expositions, however, were not able to handle extrinsic information extraction itself as a projection on a iteration invariant set, or to justify it within the context of projections algorithms, and it is this point which makes our connection useful and novel.

1.1 Purpose

The purpose of connecting EP with iterative projections algorithms, which has led to many attempts to do so, is to study its convergence behavior and to predict when it will converge. Thus, the suggestion of this paper is that a sizable convergence literature on Dykstra's algorithm with cyclic Bregman projections and alternating Bregman projections can be brought to bear on EP. Such a study will not be possible without an interpretation of EP within this context, motivating the necessity for the results in this paper.

2 Expectation Propagation

EP, first proposed in [5], defines a family of algorithms for approximate Bayesian statistical inference which exploit structure in the joint probability density function $\mathsf{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r},\boldsymbol{\theta})$ for $\mathbf{r}$ and $\boldsymbol{\theta}$. In particular, it is assumed that the $\mathsf{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r},\boldsymbol{\theta})$ factors multiplicatively

$$\mathsf{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r},\boldsymbol{\theta}) \propto \prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathsf{a},\mathbf{r}}(\boldsymbol{\theta}_\mathsf{a}), \quad \boldsymbol{\theta}_\mathsf{a} \subseteq \boldsymbol{\theta} \tag{1}$$

where the factors $\mathsf{f}_{\mathsf{a},\mathbf{r}}$ implicitly are functions of $\mathbf{r}$ and have range $[0,\infty)$ and where $\boldsymbol{\theta}_\mathsf{a}$ is a vector formed by taking a subset of the elements of $\boldsymbol{\theta}$. EP aims at iteratively approximating the joint density as the product of $\mathtt{M}$ minimal standard exponential family densities

$$\mathsf{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r},\boldsymbol{\theta}) \approx \prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{g}_{\mathsf{a},\boldsymbol{\lambda}_\mathsf{a}(\mathbf{r})}(\boldsymbol{\theta}_\mathsf{a}) \tag{2}$$

The minimal standard exponential family densities $\mathsf{g}_{\mathsf{a},\boldsymbol{\lambda}_\mathsf{a}}(\boldsymbol{\theta}_\mathsf{a})$ are the Radon Nikodym derivative of a standard exponential family measure with respect to the reference measure $\mathsf{d}\boldsymbol{\theta}_\mathsf{a}$ formed by the product of either Lebesgue or counting measures $\mathsf{d}\theta_\mathtt{i}$ for each $\theta_\mathtt{i}$ appearing in $\boldsymbol{\theta}_\mathsf{a}$. These standard exponential family densities may be parameterized in terms of a vector of real valued parameters $\boldsymbol{\lambda}_\mathsf{a}$ and sufficient statistics $\mathbf{t}_\mathsf{a}(\boldsymbol{\theta}_\mathsf{a})$ as

$$\mathsf{g}_{\mathsf{a},\boldsymbol{\lambda}_\mathsf{a}}(\boldsymbol{\theta}_\mathsf{a}) := \exp\left(\mathbf{t}_\mathsf{a}(\boldsymbol{\theta}_\mathsf{a}) \cdot \boldsymbol{\lambda}_\mathsf{a} - \psi_{\mathbf{t}_\mathsf{a}}(\boldsymbol{\lambda}_\mathsf{a})\right), \quad \psi_{\mathbf{t}_\mathsf{a}}(\boldsymbol{\lambda}_\mathsf{a}) := \log\left(\int_{\Theta_\mathsf{a}} \exp(\mathbf{t}_\mathsf{a}(\boldsymbol{\theta}_\mathsf{a}) \cdot \boldsymbol{\lambda}_\mathsf{a})\mathsf{d}\boldsymbol{\theta}_\mathsf{a}\right) \tag{3}$$

The $\mathsf{g}_{\mathsf{a},\boldsymbol{\lambda}_\mathsf{a}}$ are iteratively refined in order to approximate $\mathsf{f}_{\mathsf{a},\mathbf{r}}$ for a particular $\mathbf{r}$ by minimizing the Kullback Leibler divergence

$$\mathsf{g}_{\mathsf{a},\boldsymbol{\lambda}_\mathsf{a}} = \arg\min_{\mathsf{g}_{\mathsf{a},\boldsymbol{\lambda}_\mathsf{a}}} \mathfrak{D}\left(\mathsf{v}_\mathsf{a} \,\|\mathsf{q}\right), \quad \mathsf{v}_\mathsf{a}(\boldsymbol{\theta}) := \alpha \mathsf{f}_{\mathsf{a},\mathbf{r}}(\boldsymbol{\theta}_\mathsf{a}) \prod_{\mathsf{c}\neq\mathsf{a}} \mathsf{g}_{\mathsf{c},\boldsymbol{\lambda}_\mathsf{c}}(\boldsymbol{\theta}_\mathsf{c}), \ \mathsf{q}(\boldsymbol{\theta}) := \beta \prod_{\mathsf{c}=1}^{\mathtt{M}} \mathsf{g}_{\mathsf{c},\boldsymbol{\lambda}_\mathsf{c}}(\boldsymbol{\theta}_\mathsf{c}) \tag{4}$$

and α, β are normalization constants which ensure the densities integrate to unity. This minimization has a unique solution due to the log convexity of the Kullback Leibler distance in the second argument and the minimality of each of the representations of the standard exponential families $\mathsf{g}_{\mathsf{a},\boldsymbol{\lambda}_\mathsf{a}}$. The minima may be found by taking derivatives with respect to $\boldsymbol{\lambda}_\mathsf{a}$ to get

$$\nabla_{\boldsymbol{\lambda}_\mathsf{a}} \mathfrak{D} = \mathbb{E}_\mathsf{q}\left[\mathbf{t}_\mathsf{a}(\boldsymbol{\theta}_\mathsf{a})\right] - \mathbb{E}_{\mathsf{v}_\mathsf{a}}\left[\mathbf{t}_\mathsf{a}(\boldsymbol{\theta}_\mathsf{a})\right] \tag{5}$$

Next, another $\mathsf{g}_{\mathsf{a},\boldsymbol{\lambda}_\mathsf{a}}$ is selected to refine, usually according to some iteration order, and the algorithm continues until it converges.

The order according to which $\boldsymbol{\lambda}_\mathsf{a}$ is selected for refinement, which we call scheduling, varies in implementations. Several possibilities include *parallel scheduling* for which we update all of the parameters $\boldsymbol{\lambda}_\mathsf{a}$ in parallel, *serial scheduling* for which we update each of the parameters $\boldsymbol{\lambda}_\mathsf{a}$ in serial, and *random scheduling* when we decide which $\boldsymbol{\lambda}_\mathsf{a}$ to update by drawing $\mathtt{a}$ uniformly from the set $\{1,\ldots,\mathtt{M}\}$. We will focus on the parallel scheduling in the remainder of the article, although the results may be generalized the serial scheduling case as well.

3 Prerequisites

In this section we review background material from convex analysis concerning Bregman divergences, some associated projection algorithms based on these divergences, and the information geometry of exponential families.

3.1 Convex Analysis: Bregman Divergences

To begin, it will be useful to recall that a differentiable convex function is always lower bounded by its first order Taylor approximation. In particular, if h is a convex function of Legendre type [6] then

$$\mathsf{h}(\boldsymbol{\chi}) \geq \mathsf{h}(\boldsymbol{\varsigma}) + \nabla\mathsf{h}(\boldsymbol{\varsigma})\cdot(\boldsymbol{\chi}-\boldsymbol{\varsigma}) \tag{6}$$

with equality if and only if $\boldsymbol{\chi}=\boldsymbol{\varsigma}$. Given, then, such a convex function h, we can define a second function which has many of the properties of a notion of a distance. In particular, the *Bregman Divergence* [6], B_h associated with h is defined as

$$\mathsf{B}_\mathsf{h}(\boldsymbol{\chi},\boldsymbol{\varsigma}) := \mathsf{h}(\boldsymbol{\chi}) - \mathsf{h}(\boldsymbol{\varsigma}) - \nabla\mathsf{h}(\boldsymbol{\varsigma})\cdot(\boldsymbol{\chi}-\boldsymbol{\varsigma})$$

This function has some of the properties of a distance. In particular, we see from (6) that

$$\mathsf{B}_\mathsf{h}(\boldsymbol{\chi},\boldsymbol{\varsigma}) \geq 0 \quad \mathsf{B}_\mathsf{h}(\boldsymbol{\chi},\boldsymbol{\varsigma}) = 0 \;\Leftrightarrow\; \boldsymbol{\chi}=\boldsymbol{\varsigma}$$

On the other hand, B_h will not necessarily satisfy the triangle inequality in general and will not necessarily be symmetric.

3.2 Projection Algorithms on Convex Sets

Endowing the space now with this distance like function we are free to consider projection algorithms using B_h as the "distance" to minimize. In particular, since B_h is not in general symmetric, given a point $\boldsymbol{\chi}$ and a convex set $\mathcal{C}$ we can define two different projections of $\boldsymbol{\chi}$ onto a convex set $\mathcal{C}$ depending on which argument of B_h we minimize. We thus define the left and right projections

$$\overleftarrow{\mathbf{p}}_{\mathcal{C}}\boldsymbol{\chi} := \arg\min_{\boldsymbol{\varsigma}\in\mathcal{C}} \mathsf{B}_\mathsf{h}(\boldsymbol{\varsigma},\boldsymbol{\chi}), \quad \overrightarrow{\mathbf{p}}_{\mathcal{C}}\boldsymbol{\chi} := \arg\min_{\boldsymbol{\varsigma}\in\mathcal{C}} \mathsf{B}_\mathsf{h}(\boldsymbol{\chi},\boldsymbol{\varsigma})$$

We shall presently consider iterative projection algorithms built from one or both of these projection operators.

3.2.1 Cyclic Bregman Projections

One particular algorithm, called the *method of cyclic Bregman projections*[7], aims to solve a convex feasibility problem in which one wishes to find a point common to a finite number of convex sets $\mathcal{C}_0,\ldots,\mathcal{C}_{\mathtt{S}-1}$. In order to find such a point in their intersection $\bigcap_{\mathtt{i}=0}^{\mathtt{S}-1}\mathcal{C}_\mathtt{i}$, the algorithm iteratively projects the candidate point $\boldsymbol{\chi}^{(\mathtt{k})}$ onto one of the convex sets, choosing which convex set to project onto in a cyclic manner. Of course, there are two algorithms which fall into this category, one if the projections are done in the first argument of the divergence

$$\boldsymbol{\chi}^{(\mathtt{k}+1)} := \overleftarrow{\mathbf{p}}_{\mathcal{C}_{\mathtt{k} \bmod \mathtt{S}}}\boldsymbol{\chi}^{(\mathtt{k})}$$

and another if the projections are done in the second argument of the divergence

$$\boldsymbol{\chi}^{(\mathtt{k}+1)} := \overrightarrow{\mathbf{p}}_{\mathcal{C}_{\mathtt{k} \bmod \mathtt{S}}}\boldsymbol{\chi}^{(\mathtt{k})}$$

3.2.2 Dykstra's Algorithm

A close variant of the method of cyclic Bregman projections, often called *Dykstra's algorithm* [6, 8, 9], adds some extra processing between the projections. In particular, denoting by h^* the convex conjugate of h, Dykstra's algorithm is the iteration

$$\boldsymbol{\chi}^{(\mathtt{k}+1)} := \overleftarrow{\mathbf{p}}_{\mathcal{C}_{\mathtt{k} \bmod \mathtt{S}}} \nabla \mathsf{h}^* \left(\nabla \mathsf{h}(\boldsymbol{\chi}^{(\mathtt{k})}) + \boldsymbol{\tau}^{(\mathtt{k}+1-\mathtt{S})} \right), \quad \boldsymbol{\tau}^{(\mathtt{k}+1)} := \nabla \mathsf{h}(\boldsymbol{\chi}^{(\mathtt{k})}) + \boldsymbol{\tau}^{(\mathtt{k}+1-\mathtt{S})} - \nabla \mathsf{h}(\boldsymbol{\chi}^{(\mathtt{k}+1)})$$

where we initialize $\boldsymbol{\tau}^{(-\mathtt{S}+1)}, \ldots, \boldsymbol{\tau}^{(0)} = \mathbf{0}$. This algorithm, under some assumptions, can be shown [6] to solve the best approximation problem, in which one is seeking the point in $\mathcal{C} := \bigcap_{\mathtt{i}=0}^{\mathtt{S}-1} \mathcal{C}_{\mathtt{i}}$ which minimizes the Bregman divergence B_{h} in the first argument from the initial point $\boldsymbol{\chi}^{(0)}$. That is, the sequence $\{\boldsymbol{\chi}^{(\mathtt{k})}\}$ converges to $\overleftarrow{\mathbf{p}}_{\mathcal{C}} \boldsymbol{\chi}^{(0)}$.

3.2.3 Alternating Bregman Projections

In another situation, it is desirable to find points in two different convex sets $\mathcal{P}$ and $\mathcal{Q}$ which minimize the Bregman divergence B_{h} between these two sets. The projection algorithm that is often employed in this case is the *method of alternating projections* [10, 11, 12] which may be described via the iteration

$$\boldsymbol{\chi}^{(\mathtt{k})} := \overleftarrow{\mathbf{p}}_{\mathcal{P}} \boldsymbol{\varsigma}^{(\mathtt{k})}, \quad \boldsymbol{\varsigma}^{(\mathtt{k}+1)} := \overrightarrow{\mathbf{p}}_{\mathcal{Q}} \boldsymbol{\chi}^{(\mathtt{k})}$$

3.3 Information Geometry of Exponential Families

The usefulness of projection algorithms built from Bregman divergences arises in the context of EP through its use of exponential families of probability distributions. There are two natural parameter spaces for exponential families related through the Legendre transformation (Fenchel convex conjugation) between the entropy and the log partition function potential functions [13, 14, 15]. These two potential functions, the entropy and the log partition function, yield the Kullback Leibler divergence, also known as the relative entropy, as their Bregman divergence. Since the EP algorithm is defined as an iterative algorithm minimizing the Kullback Leibler divergences between several iteration-varying measures, we can see the relevance between projection algorithms using Bregman divergences and EP. We presently attempt to review some of the relevant material from the information geometry of exponential families that is relevant for the result about EP that follows.

Exponential families are collections of probability measures which have probability densities which can be written in the form

$$\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \exp\left(\mathbf{t}(\boldsymbol{\theta}) \cdot \boldsymbol{\lambda} - \psi_{\mathbf{t}}(\boldsymbol{\lambda})\right), \quad \psi_{\mathbf{t}}(\boldsymbol{\lambda}) := \log\left(\int_{\Theta} \exp\left(\mathbf{t}(\boldsymbol{\theta}) \cdot \boldsymbol{\lambda} - \psi_{\mathbf{t}}(\boldsymbol{\lambda})\right) \mathsf{d}\boldsymbol{\theta} \right) \tag{7}$$

The sufficient statistics functions $\mathbf{t}(\cdot)$ choose the exponential family, and the parameters $\boldsymbol{\lambda}$ choose the probability distribution within the family. The function $\psi_{\mathbf{t}}(\boldsymbol{\lambda})$, called the log partition function by statistical physicists, is defined so that the probability density integrates/sums to one. Some examples of exponential families include the Gaussian distributions, multivariate Gaussian distributions, all probability distributions on a finite set, the exponential distribution, any joint distribution for independent exponential family random variables, etc. To give concrete meaning, note that identifying $\Theta := \mathbb{R}$, $\mathbf{t}(\theta) := [\theta, \theta^2]^T$, gives the family of Gaussian PDFs. The particular Gaussian is then selected by choosing the mean m and variance σ^2, through the relation $\boldsymbol{\lambda} := [m/\sigma^2, -1/(2\sigma^2)]^T$. We will assume that the chosen sufficient statistics are such that the representation is minimal.

In equation (7) we have used the log coordinate $\boldsymbol{\lambda}$ parametrization of the exponential family. We may also select a particular distribution within a given exponential family via the expectation coordinates

$$\boldsymbol{\eta} := \int_{\Theta} \mathbf{t}(\boldsymbol{\theta}) \exp\left(\mathbf{t}(\boldsymbol{\theta}) \cdot \boldsymbol{\lambda} - \psi_{\mathbf{t}}(\boldsymbol{\lambda})\right) \mathsf{d}\boldsymbol{\theta}$$

This map between $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ is one to one, and may be interpreted to be the gradient of the log partition function

$$\boldsymbol{\eta} = \nabla_{\boldsymbol{\lambda}} \psi_{\mathbf{t}}(\boldsymbol{\lambda})$$

This relation indicates a Legendre transformation connection between $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$. In particular, due to the convexity of the log partition partition function, we can form its Fenchel conjugate as

$$\mathsf{h}(\boldsymbol{\eta}) := -\inf_{\boldsymbol{\lambda}_1} \{\psi_{\mathbf{t}}(\boldsymbol{\lambda}_1) - \boldsymbol{\eta} \cdot \boldsymbol{\lambda}_1\} = \psi_{\mathbf{t}}(\boldsymbol{\lambda}) - \boldsymbol{\eta} \cdot \boldsymbol{\lambda}$$

This function can be recognized as the negative of the Shannon entropy [13].

The relationship between the log and expectation coordinates and the entropy and partition function may be summarized in the following convenient inequality: for two different probability distributions p and q, p having expectation coordinates $\boldsymbol{\eta}_p$ and the q with log coordinates $\boldsymbol{\lambda}_q$, we have

$$\psi(\boldsymbol{\lambda}_q) + \mathsf{h}(\boldsymbol{\eta}_p) = \boldsymbol{\eta}_p \cdot \boldsymbol{\lambda}_q + \mathfrak{D}(p||q) \tag{8}$$

where $\mathfrak{D}$ is the Kullback Leibler distance (a.k.a. the relative entropy).

We now have two convex functions, the partition function in terms of the log coordinates $\boldsymbol{\lambda}$, and the negative of the Shannon entropy in terms of the expectation coordinates $\boldsymbol{\eta}$, which we may build Bregman divergences from. In fact, we may directly infer from the relationship (8) that the associated Bregman divergence is the relative entropy. From here on, we choose to work with expectation coordinates, so we generate the Bregman divergence generated from the negative entropy. To keep our notation close to that of the generic alternating Bregman projection algorithms, we will denote the convex conjugate of h with h^* (as opposed to ψ) from here on out.

The expectation maximization algorithm for statistical inference in the presence of nuisance parameters (which is different from expectation *propagation*) may be interpreted, for instance, as an alternating projection algorithm using the Kullback Leibler divergence at the Bregman divergence [10, 13]. [16] and [13] discusses some properties and well-posedness of projections when the Bregman divergence is the Kullback Leibler divergence.

4 EP as Iterated Projections

For the convergence framework that we are about to apply to EP to be valid, we will need to make a rather benign extra assumption about the structure of the true a posteriori distribution $\mathsf{p}_{\boldsymbol{\theta}|\mathbf{r}}(\boldsymbol{\theta}|\mathbf{r})$ for the parameters $\boldsymbol{\theta}$ given the observations $\mathbf{r}$. In particular, we must assume that it is a exponential family distribution, say with sufficient statistics $\mathbf{k}(\boldsymbol{\theta})$, so that

$$\mathsf{p}_{\boldsymbol{\theta}|\mathbf{r}}(\boldsymbol{\theta}|\mathbf{r}) \in \{\exp\left(\boldsymbol{\kappa} \cdot \mathbf{k}(\boldsymbol{\theta}) - \psi_{\mathbf{k}}(\boldsymbol{\kappa})\right)\}$$

for some particular parameter $\boldsymbol{\kappa} = \boldsymbol{\kappa}(\mathbf{r})$. Furthermore, we need to make the extra assumption that the approximating standard exponential family $\mathsf{g}(\boldsymbol{\theta})$ with sufficient statistics $\mathbf{t}$ is a subfamily of the standard exponential family with sufficient statistics $\mathbf{k}$, so that

$$\{\exp\left(\mathbf{t}(\boldsymbol{\theta}) \cdot \boldsymbol{\lambda} - \psi_{\mathbf{t}}(\boldsymbol{\lambda})\right)\} \subseteq \{\exp\left(\mathbf{k}(\boldsymbol{\theta}) \cdot \boldsymbol{\kappa} - \psi_{\mathbf{k}}(\boldsymbol{\kappa})\right)\}$$

These conditions will hold, for instance, if the parameters $\boldsymbol{\theta}$ are drawn from a finite set. It is important to note that this condition does *not* require that the family of marginal distributions of the true a posteriori density $\mathsf{p}_{\boldsymbol{\theta}|\mathbf{r}}(\boldsymbol{\theta}|\mathbf{r})$ equal the family with sufficient statistics $\mathbf{t}$ (as would be the case with belief propagation). Rather, the two families (that of the approximating distribution and that of the true a posteriori distributions) just need to be (usually different) subfamilies of the same family of exponential densities. Furthermore, we need not work with $\mathbf{k}$ directly in an implementation, so it can be very high dimensional and unwieldy. As we shall see when presenting the main theorem, its introduction is just an artifact of the proof of the interpretation of EP as a Dykstra-like projection algorithm between two iteration invariant sets.

Consider now exponential family distributions on $\Theta^{\mathtt{M}}$, that is probability distributions on the product space of $\mathtt{M}$ copies of the original parameter space $\boldsymbol{\theta}$. We will denote an element of this space by $\mathbf{x} := \left[\mathbf{x}^1, \ldots, \mathbf{x}^{\mathtt{M}}\right]$, so that we are considering probability distributions on $\mathbf{x}$ (the superscript notation is to avoid confusion with the subscript notation which meant subsets of the original vector instead of different vectors). In particular we will be interested in the set $\mathcal{B}$

of exponential family distributions on $\mathbf{x}$ with sufficient statistics $\mathbf{s}(\mathbf{x}) := \left[\mathbf{k}(\mathbf{x}^1)^T, \cdot, \mathbf{k}(\mathbf{x}^{\mathtt{M}})^T\right]^T$ (and reference measure which the product of $\mathtt{M}$ copies of the original reference measure). It is subsets of this family which we will consider in our discussion of EP algorithms as projection algorithms.

Of particular interest will be the set $\mathcal{Q}$ of distributions from $\mathcal{B}$ which give $\mathbb{P}\left[\mathbf{x}^1 = \cdots = \mathbf{x}^{\mathtt{M}}\right] = 1$, so that the probability distributions in $\mathcal{Q}$ all only have probability in the subspace of $\Theta^{\mathtt{M}}$ for which each of the "copied" random variables $\mathbf{x}^1$ are equal almost surely.

$$\mathcal{Q} := \left\{\mathsf{b} \in \mathcal{B} | \mathbb{P}_{\mathsf{b}}\left[\mathbf{x}^1 = \cdots = \mathbf{x}^{\mathtt{M}}\right] = 1\right\} \tag{9}$$

We will also be interested in the set $\mathcal{P}$ of exponential family distributions with sufficient statistics

$$\hat{\mathbf{t}}(\mathbf{x}) := \left[\mathbf{t}(\mathbf{x}^1), \ldots, \mathbf{t}(\mathbf{x}^{\mathtt{M}})\right]$$

where $\mathbf{t}$ is the vector concatenation of all linearly independent functions that are elements of any of the $\mathbf{t}_{\mathsf{a}}(\boldsymbol{\theta}_{\mathsf{a}})$, so that the exponential family distributions with sufficient statistics $\mathbf{t}_{\mathsf{a}}$ comprise the family of distributions within which EP is trying to choose g to approximate the true a posteriori distribution as in (2).

$$\mathcal{P} := \left\{\mathsf{b} | \mathsf{b} = \exp\left(\boldsymbol{\lambda} \cdot \hat{\mathbf{t}}(\mathbf{x}) - \psi_{\hat{\mathbf{t}}}(\boldsymbol{\lambda})\right),\ \boldsymbol{\lambda} \in \mathbb{R}^{\mathtt{MV}}\right\} \tag{10}$$

Finally, we will be working with the negative entropy function h on $\mathcal{B}$ written as a function of the expectation parameters $\boldsymbol{\eta} := \mathbb{E}\left[\mathbf{s}(\mathbf{x})\right]$.

Oftentimes, the desired result of applying EP is an approximating density whose expected values of the sufficient statistics match that of the true a posteriori density, so that

$$\mathbb{E}_{\mathsf{g}}\left[\mathbf{t}(\boldsymbol{\theta})\right] = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{r}}\left[\mathbf{t}(\boldsymbol{\theta})\right] \tag{11}$$

The next proposition connects this ideal solution as a composition of two Bregman projections using the Kullback Leibler divergence.

Prop. 1 (Optimal Solution as Two Projections)**:** Consider the Bregman divergence generated using the negative entropy h function on $\mathcal{B}$ written in terms of $\mathbb{E}[\mathbf{s}(\mathbf{x})]$ as its convex function. Then, the Bregman projection defined by

$$\overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \overleftarrow{\mathbf{p}}_{\mathcal{Q}} \left(\frac{\int_{\Theta^{\mathtt{M}}} \mathbf{s}(\mathbf{x}) \prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathsf{a}}(\mathbf{x}^{\mathsf{a}}) d\mathbf{x}}{\int_{\Theta^{\mathtt{M}}} \prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathsf{a}}(\mathbf{x}^{\mathsf{a}}) d\mathbf{x}}\right) \tag{12}$$

yields expectation parameters $\mathbb{E}[\hat{\mathbf{t}}]$ corresponding to a probability distribution $\mathsf{g}(\mathbf{x}) \in \mathcal{B}$ which satisfies

$$\mathbb{E}_{\mathsf{g}}[\mathbf{t}(\mathbf{x}^{\mathsf{a}})] = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{r}}\left[\mathbf{t}(\boldsymbol{\theta})\right] \quad \forall \mathsf{a} \in \{1, \ldots, \mathtt{M}\}$$

Proof: The first projection $\overleftarrow{\mathbf{p}}_{\mathcal{Q}}\left(\prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathsf{a}}(\mathbf{x}^{\mathsf{a}})\right)$ yields as its result the expectation coordinates of a probability distribution

$$\frac{\prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathsf{a}}(\mathbf{x}^1)\delta(\mathbf{x}^{\mathsf{a}} - \mathbf{x}^1)}{\int_{\Theta} \prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathsf{a}}(\mathbf{x}^1) d\mathbf{x}^1} =: \mathsf{p}_{\boldsymbol{\theta}|\mathbf{r}}(\mathbf{x}^1|\mathbf{r}) \prod_{\mathsf{c}=2}^{\mathtt{M}} \delta(\mathbf{x}^1 - \mathbf{x}^{\mathsf{c}})$$

where δ is (either the Kronecker or Dirac) point mass distribution, since such a distribution attains a Kullback Leibler divergence of zero (which is the global minimum value). The second projection $\overrightarrow{\mathbf{p}}_{\mathcal{P}}$ then, calculates the expectation of sufficient statistics (as we showed by taking derivatives in Section 2) to choose the approximating distribution (and thus result of the projection). Thus, defining

$$\mathsf{f} := \overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \overleftarrow{\mathbf{p}}_{\mathcal{Q}} \left(\prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathsf{a}}(\mathbf{x}^{\mathsf{a}})\right)$$

implies $\mathbb{E}_{\mathsf{f}}[\mathbf{t}(\mathbf{x}^{\mathsf{a}})] = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{r}}[\mathbf{t}(\boldsymbol{\theta})]$ for any $\mathsf{a} \in \{1, \ldots, \mathtt{M}\}$. ■

Of course, EP does not perform the projection (12) directly, but if it is to be a method for approximately satisfying (11), as suggested in [17] among many other places, then it may be considered an iterative projection method attempting to approximately solve the projection problem (12).

Thm. 1 (EP as Information Projections)**:** The EP algorithm under parallel scheduling is equivalent to the following Dykstra-like iterated Bregman projections algorithm

$$\begin{aligned}
\boldsymbol{\rho}_0, \boldsymbol{\tau}_0 &= \mathbf{0}, \quad \boldsymbol{\chi}_0 = \frac{\int_{\Theta^{\mathtt{M}}} \mathbf{s}(\mathbf{x}) \prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathsf{a}}(\mathbf{x}^{\mathsf{a}}) d\mathbf{x}}{\int_{\Theta^{\mathtt{M}}} \prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathsf{a}}(\mathbf{x}^{\mathsf{a}}) d\mathbf{x}}, \quad \mathsf{k} \in \{0, 1, \ldots, \} && (13) \\
\varsigma_{\mathsf{k}} &:= \overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \nabla \mathsf{h}^* \left(\nabla \mathsf{h}(\boldsymbol{\chi}_{\mathsf{k}}) + \boldsymbol{\tau}_{\mathsf{k}} \right), \quad \boldsymbol{\tau}_{\mathsf{k}+1} := \nabla \mathsf{h}(\boldsymbol{\chi}_{\mathsf{k}}) + \boldsymbol{\tau}_{\mathsf{k}} - \nabla \mathsf{h}(\varsigma_{\mathsf{k}}) && (14) \\
\boldsymbol{\chi}_{\mathsf{k}+1} &:= \overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \overleftarrow{\mathbf{p}}_{\mathcal{Q}} \circ \nabla \mathsf{h}^* \left(\nabla \mathsf{h}(\varsigma_{\mathsf{k}}) + \boldsymbol{\rho}_{\mathsf{k}} \right), \quad \boldsymbol{\rho}_{\mathsf{k}+1} := \nabla \mathsf{h}(\varsigma_{\mathsf{k}}) + \boldsymbol{\rho}_{\mathsf{k}} - \nabla \mathsf{h}(\boldsymbol{\chi}_{\mathsf{k}+1}) && (15)
\end{aligned}$$

where the Bregman divergence is built from the convex function h which is the negative entropy function on $\mathcal{B}$ written in terms of $\mathbb{E}[\mathbf{s}(\mathbf{x})]$, and $\mathcal{P}$ $\mathcal{Q}$ are the convex and log convex sets defined in (10) and (9). In this equivalence $\nabla \mathsf{h}(\boldsymbol{\chi}_{\mathsf{k}}) + \boldsymbol{\tau}_{\mathsf{k}}$ are the log coordinates of the density $\prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{v}_{\mathsf{a}}(\mathbf{x}^{\mathsf{a}})$ and $\boldsymbol{\chi}_{\mathsf{k}}$ are the expectation coordinates of the density $\prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{q}(\mathbf{x}^{\mathsf{a}})$ with v_{a} and q being defined as in (4).

Proof: Consider the first iteration,

$$\varsigma_0 := \overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \nabla \mathsf{h}^* \left(\nabla \mathsf{h}(\boldsymbol{\chi}_0) + \boldsymbol{\tau}_0 \right) = \overrightarrow{\mathbf{p}}_{\mathcal{P}} \boldsymbol{\chi}_0$$

which simultaneously updates g_{a} for all a according to equation (4). This is followed by

$$\boldsymbol{\chi}_1 := \overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \overleftarrow{\mathbf{p}}_{\mathcal{Q}} \circ \nabla \mathsf{h}^* \left(\nabla \mathsf{h}(\varsigma_0) + \boldsymbol{\rho}_0 \right) = \overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \overleftarrow{\mathbf{p}}_{\mathcal{Q}} \circ \overrightarrow{\mathbf{p}}_{\mathcal{P}} \boldsymbol{\chi}_0$$

The projection onto $\mathcal{Q}$ forms the product over a of the new g_{a}s, and the next projection onto $\mathcal{P}$ forms $\mathtt{M}$ copies of that product. Next, since

$$\boldsymbol{\tau}_1 := \nabla \mathsf{h}(\boldsymbol{\chi}_0) + \boldsymbol{\tau}_0 - \nabla \mathsf{h}(\varsigma_0) = \nabla \mathsf{h}(\boldsymbol{\chi}_0) - \nabla \mathsf{h}(\overrightarrow{\mathbf{p}}_{\mathcal{P}} \boldsymbol{\chi}_0)$$

we have

$$\varsigma_1 := \overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \nabla \mathsf{h}^* \left(\nabla \mathsf{h}(\boldsymbol{\chi}_1) + \boldsymbol{\tau}_1 \right) = \overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \nabla \mathsf{h}^* \left(\nabla \mathsf{h}(\overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \overleftarrow{\mathbf{p}}_{\mathcal{Q}} \circ \overrightarrow{\mathbf{p}}_{\mathcal{P}} \boldsymbol{\chi}_0) + \nabla \mathsf{h}(\boldsymbol{\chi}_0) - \nabla \mathsf{h}(\overrightarrow{\mathbf{p}}_{\mathcal{P}} \boldsymbol{\chi}_0) \right)$$

Furthermore, the addition (subtraction) of $\nabla \mathsf{h}$s is equivalent to adding the log coordinates of the associated densities, which is equivalent to multiplying (dividing) the associated densities. This then gives that $\nabla \mathsf{h}(\overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \overleftarrow{\mathbf{p}}_{\mathcal{Q}} \circ \overrightarrow{\mathbf{p}}_{\mathcal{P}} \boldsymbol{\chi}_0) + \nabla \mathsf{h}(\boldsymbol{\chi}_0) - \nabla \mathsf{h}(\overrightarrow{\mathbf{p}}_{\mathcal{P}} \boldsymbol{\chi}_0)$ is equal to the log coordinates of the density $\prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{v}_{\mathsf{a}}(\mathbf{x}^{\mathsf{a}})$, and thus the ς_1 is the expectation coordinates of the density $\prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{g}_{\mathsf{a}}^{(1)}(\mathbf{x}_{\mathsf{a}}^{\mathsf{a}}) \prod_{\mathsf{c} \neq \mathsf{a}} \mathsf{g}_{\mathsf{c}}^{(0)}(\mathbf{x}_{\mathsf{c}}^{\mathsf{a}})$, where the superscript indicates iteration number. Via the same justification, $-\boldsymbol{\rho}_1$ must be the log coordinates of the density $\prod_{\mathsf{a}=1}^{\mathtt{M}} \prod_{\mathsf{c} \neq \mathsf{a}} \mathsf{g}_{\mathsf{c}}^{(0)}(\mathbf{x}_{\mathsf{c}}^{\mathsf{a}})$, so adding $\boldsymbol{\rho}_1$ to $\nabla \mathsf{h}(\varsigma_1)$ forms the log coordinates of $\prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{g}_{\mathsf{a}}^{(1)}(\mathbf{x}^{\mathsf{a}})$. The next projection then makes $\boldsymbol{\chi}_1$ the expectation coordinates of $\prod_{\mathsf{c}=1}^{\mathtt{M}} \prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{g}_{\mathsf{a}}^{(1)}(\mathbf{x}^{\mathsf{c}})$, and we may cycle back through the same chain of arguments to form the induction step for the subsequent iterations. ■

Note that there are several marked differences between EP and Dykstra's algorithm. To begin with, Dykstra's algorithm with cyclic Bregman projections is built with only left proximity operators, and it is somewhat unusual for one of the projections to not have any preprocessing. One can consider alternating projections with Dykstra preprocessing ([11] did so for Hilbert spaces with orthogonal projections instead of Bregman projections), but one must use different preprocessing based on the direction of the projection to follow it. The preprocessing used in EP is the preprocessing for a left projection, but it is followed by a right projection. Still, the similarity between the two algorithms is striking, and it suggests that convergence results might be developed via this analogy.

References

[1] M. Moher and T. A. Gulliver, "Cross-entropy and iterative decoding.," *IEEE Trans. Inform. Theory*, vol. 44, pp. 3097–3104, Nov. 1998.

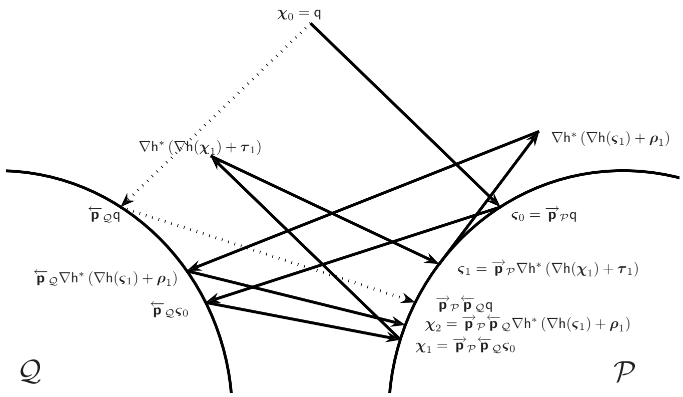
$\chi_0 = q$
$\nabla h^* (\nabla h(\chi_1) + \tau_1)$
$\nabla h^* (\nabla h(\varsigma_1) + \rho_1)$
$\overleftarrow{p}_{\mathcal{Q}} q$
$\varsigma_0 = \overrightarrow{p}_{\mathcal{P}} q$
$\varsigma_1 = \overrightarrow{p}_{\mathcal{P}} \nabla h^* (\nabla h(\chi_1) + \tau_1)$
$\overleftarrow{p}_{\mathcal{Q}} \nabla h^* (\nabla h(\varsigma_1) + \rho_1)$
$\overrightarrow{p}_{\mathcal{P}} \overleftarrow{p}_{\mathcal{Q}} q$
$\overleftarrow{p}_{\mathcal{Q}} \varsigma_0$
$\chi_2 = \overrightarrow{p}_{\mathcal{P}} \overleftarrow{p}_{\mathcal{Q}} \nabla h^* (\nabla h(\varsigma_1) + \rho_1)$
$\chi_1 = \overrightarrow{p}_{\mathcal{P}} \overleftarrow{p}_{\mathcal{Q}} \varsigma_0$
$\mathcal{Q}$
$\mathcal{P}$

Figure 1: EP (solid arrows), and the composite projection problem it iteratively solves (dotted arrows), but with $\|\|_2^2$ as the Bregman divergence and different sets.

[2] B. Muquet, P. Duhamel, and M. de Courville, "Geometrical interpretations of iterative 'turbo' decoding," in *Proceedings ISIT*, June 2002.

[3] S. Ikeda, T. Tanaka, and S. Amari, "Information geometry of turbo and low-density parity-check codes," *IEEE Trans. Inform. Theory*, vol. 50, pp. 1097 – 1114, June 2004.

[4] S. Ikeda, T. Tanaka, and S. Amari, "Stochastic reasoning, free energy and information geometry," *Neural Computation*, pp. 1779–1810, 2004.

[5] T. P. Minka, *A Family of Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Massachusetts Institute of Technology, 2001.

[6] H.H. Bauschke and A.S. Lewis, "Dykstra's algorithm with Bregman projections: a convergence proof," *Optimization*, , no. 48, pp. 409–427, 2000.

[7] L.G. Gubin, B.T. Polyak, and E.V. Raik, "The method of projections for finding the common point of convex sets," *USSR J. Comp. Math. and Math. Phys.*, vol. 7, no. 6, pp. 1211–1228, 1967.

[8] L.M. Bregman, Y. Censor, and S. Reich, "Dykstra's Algorithm as the Nonlinear Extension of Bregman's Optimization Method," *J. Conv. Anal.*, vol. 6, no. 2, pp. 319–333, 1999.

[9] R.L. Dykstra, "An iterative procedure for obtaining *i*-projections onto the intersection of convex sets," *Annals of Prob.*, vol. 13, no. 3, pp. 975–984, Aug. 1985.

[10] I. Csiszár and G. Tusnády, "Information geometry and alterating minimization procedures," *Statistics and Decisions, Supplement Issue*, pp. 205–237, 1984.

[11] H.H. Bauschke and J.M. Borwein, "Dykstra's alternating projection algorithm for two sets," *J. Approx. Theory*, , no. 79(3), pp. 418–443, 1994.

[12] H.H. Bauschke, P. L. Combettes, and D. Noll, "Joint minimization with alternating bregman proximity operators," http://mip.ups-tlse.fr/ noll/PAPERS/heinz.pdf.

[13] S. Amari, *Methods of Information Geometry*, vol. 191, AMS Translations of Mathematical Monographs, 2004.

[14] Lawrence D. Brown, *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, Institute of Mathematical Statistics, 1986.

[15] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," Tech. Rep., Department of Statistics, University of California, Berkeley, 2003.

[16] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1474–1490, June 2003.

[17] T. Minka, "Divergence measures and message passing," Tech. Rep. MSR-TR-2005-173, Microsoft Research, 2005.