Belief Propagation, Information Projections, and Dykstra’s Algorithm

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

1. Some Convex Programming
   (a) Bregman Divergence
   (b) Bregman Projections & Examples
   (c) Bregman Projections Algorithms: Alternating Bregman Projections & Dykstra’s Algorithm with Cyclic Bregman Projections

2. Belief Propagation
   (a) Definition
   (b) Applications (efficient decoders for near optimal codes)
   (c) Open Problems: Convergence and Performance conditions

3. Belief Propagation as a Hybrid Dykstra’s Algorithm
   (a) New Result: Euclidean BP Always Converges
   (b) New Characterization of Good Behavior of regular BP for Cyclic Factorings
Bregman Divergence

Convex function lower bounded by its 1st order Taylor series

\[ f(r) \geq f(q) + \langle \nabla f(q), r - q \rangle \]

for all \( r \in D \)

- \( f \) strictly convex, then strict inequality unless \( r = q \). Can then use this to construct the Bregman divergence

\[ D_f(r, q) \overset{\triangle}{=} f(r) - f(q) - \langle \nabla f(q), r - q \rangle \geq 0 \]

which vanishes \( \Leftrightarrow r = q \).

- Need not be symmetric, i.e. in general \( D_f(q, r) \neq D_f(r, q) \).

- Need not satisfy triangle inequality. (only happens in special cases)
Bregman Divergence, cont’d

- Given convex $f(q)$, have **convex conjugate**

$$f^*(\theta) = \sup_q \left( \langle q, \theta \rangle - f(q) \right).$$

- $f^*(\theta)$ is convex, and (if $f$ is of Legendre type) the gradients $\nabla f(q)$ and $\nabla f^*(\theta)$ are inverse maps to each other \([1]\),

$$\nabla f^*(\nabla f(q)) = q, \quad \nabla f(\nabla f^*(\theta)) = \theta$$

- Conjugation switches arguments

$$D_{f^*}(\nabla f(r), \nabla f(q)) = D_f(q, r)$$

- **Example: Euclidean** $f(q) = \frac{1}{2}\|q\|_2^2 = f^*(q)$, $D_f(r, q) = \frac{1}{2}\|r - q\|_2^2$
**Bregman Divergence, cont’d**

- **Example: KL Divergence** $f : \mathcal{D} \rightarrow \mathbb{R}$, $\mathcal{D} = \left\{ q \geq 0 \bigg| \sum_{i=1}^{N} q_i \leq 1 \right\}$ the negative Shannon entropy

  $$f(q) = \sum_{i=1}^{N} q_i \log(q_i) + \left( 1 - \sum_{i=1}^{N} q_i \right) \log \left( 1 - \sum_{i=1}^{N} q_i \right) = h(q)$$

  the partition function

  $$f^*(\theta) = \log(1 + \| \exp(\theta) \|_1) = \psi(\theta)$$

  with domain $\mathcal{D}^* = \mathbb{R}^N_e$
Bregman Projections

Because of asymmetry given a Bregman divergence have two notions of projections \([2, 3]\)

**Left Projection:** \(C \subset D, \text{ convex}\)

\[
\overrightarrow{P}_C q := \arg \min_{r \in C} D_f(r, q)
\]

**Right Projection:** \(P^* \subset D^*, \text{ convex, } P = \nabla f^*(P^*)\).

\[
\overrightarrow{P}_P q := \arg \min_{r \in P} D_f(q, r)
\]

Right projection with \(D_f\) can also be considered as left projection with \(D_{f^*}\).
Bregman Projections Algorithms

• **Alternating Bregman Projections** [4, 5, 6]

\[ \chi^{(k)} := \overrightarrow{P_P} \varsigma^{(k)}, \quad \varsigma^{(k+1)} := \overrightarrow{P_Q} \chi^{(k)} \]

\( P \subset \mathcal{D}, \ Q^* \subset \mathcal{D}^* \) convex. Finds points \( p \in P \) and \( q \in Q \) which obtain
\[ \inf_{p \in P, \ q \in Q} D_f(p, q). \]

• **Dykstra’s Algorithm with Cyclic Bregman Projections** [7, 8, 9]

\[ \chi^{(k+1)} := \overrightarrow{P_{C_{k \mod S}}} \nabla f^* \left( \nabla f(\chi^{(k)}) + \tau^{(k+1-S)} \right) \]
\[ \tau^{(k+1)} := \nabla f(\chi^{(k)}) + \tau^{(k+1-S)} - \nabla f(\chi^{(k+1)}) \]

with \( \tau^{(-S+1)}, \ldots, \tau^{(0)} = 0 \). Under some (benign) assumptions, [7] solves best approximation problem,
\[ \arg \min_{\chi \in \bigcap_{i=0}^{s-1} C_i} D_f(\chi, \chi^0) \]

• may also choose set to project on randomly [10, 11].
Belief Propagation

- aim to deduce $M$ bits $\mathbf{x} = [x_1, \ldots, x_M]$ based on observation of $\mathbf{y}$ and likelihood $p(\mathbf{y}|\mathbf{x})$.

$$p(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^{K} g_k(\mathbf{x})$$

$g_k$ depends on a subset of $\mathbf{x}$.

$$m_{x_i \rightarrow g_k}(x_i) = \beta_k \prod_{\ell=1}^{K} m_{g_{\ell} \rightarrow x_i}(x_i) \quad k = 1, 2, \ldots, K;$$

$$m_{g_k \rightarrow x_i}(x_i) = \alpha_i \sum_{x_{\ell}: \ell \neq i} g_k(\mathbf{x}) \prod_{n=1}^{M} m_{x_n \rightarrow g_k}(x_n), \quad i = 1, 2, \ldots, M;$$

Figure 1: A factor graph.
Belief Propagation: Applications & Problems

Applications:

• practical decoding of near channel capacity achieving codes (LDPC and turbo codes) for BEC, BSC, AWGN

• Lossless compression

• practical decoding of Slepian Wolf distributed lossless source codes.

• practical decoding of codes for the MAC

Known Problems:

• Doesn’t always converge. Can have pathological dynamics behavior.

• When it does converge, convergent points need not be “near-marginals”
Belief Propagation as a Hybrid Projections Algorithm

- Make $K$ indep. copies of $x$, forming $\bar{x} := x^1, x^2, \ldots, x^K$.
- $q \in \mathcal{D}, \theta \in \mathcal{D}^*$ parameterize the set of PMFs (subset of $N = 2^{KM} - 1$ dimensional space.)
- $f$ the negative Shannon entropy.
- $Q \subset \mathcal{D}$ the set of $q$ such that all $K$ copies are equal:
  \[ Q := \{ q | \mathbb{P}_q[x^1 = \ldots = x^K] = 1 \} \]
- $\mathcal{P}^* \subset \mathcal{D}^*$ the set of log coordinates dual to the set of product distributions over all bits
  \[ \mathcal{P} := \left\{ q(x^1, \ldots, x^K) = \prod_{k=1}^{K} \prod_{m=1}^{M} q(x_m^k) \right\} \]

Initial point is based on the factoring:
\[ \chi_0 = \prod_{k=1}^{K} g_k(x^k) \]
Belief Propagation as a Hybrid Projections Algorithm

• Desired solution (exact marginals) is two step projection

\[ \overrightarrow{P}_P \circ \overleftarrow{P}_Q \chi_0 \]

• Key Result: Belief Propagation is the Dykstra-like algorithm

\[ \varsigma_k := \overrightarrow{P}_P \circ \nabla f^* \left( \nabla f(\chi_k) + \tau_k \right) \]  \hspace{1cm} (1)

\[ \tau_{k+1} := \nabla f(\chi_k) + \tau_k - \nabla f(\varsigma_k) \]

\[ \chi_{k+1} := \overrightarrow{P}_P \circ \overleftarrow{P}_Q \circ \nabla f^* \left( \nabla f(\varsigma_k) + \rho_k \right) \]  \hspace{1cm} (2)

\[ \rho_{k+1} := \nabla f(\varsigma_k) + \rho_k - \nabla f(\chi_{k+1}) \]

with \( f \) the negative Shannon entropy (i.e. KL projections) and \( \rho_0, \tau_0 = 0 \).

• Hybrid between alternating Bregman projections and Dykstra’s algorithm with cyclic Bregman projections.
Belief Propagation as a Hybrid Projections Algorithm

\[ \chi_0 = q \]

\[ \varsigma_0 = p \q q \]

\[ \nabla h^* (\nabla h(\chi_1) + \tau_1) \]

\[ \nabla h^* (\nabla h(\varsigma_1) + \rho_1) \]

\[ \chi_1 = p \q \nabla h^* (\nabla h(\varsigma_1) + \rho_1) \]

\[ \nabla h^* (\nabla h(\chi_1) + \tau_1) \]

\[ \varsigma_1 = p \p \nabla h^* (\nabla h(\chi_1) + \tau_1) \]

\[ \chi_2 = p \p p \q \nabla h^* (\nabla h(\varsigma_1) + \rho_1) \]

\[ \nabla h^* (\nabla h(\varsigma_1) + \rho_1) \]

\[ \varsigma_0 = p \p q \]

\[ \chi_0 = q \]

\[ \varsigma_0 = p \q q \]

\[ \nabla h^* (\nabla h(\chi_1) + \tau_1) \]

\[ \nabla h^* (\nabla h(\varsigma_1) + \rho_1) \]

\[ \chi_1 = p \q \nabla h^* (\nabla h(\varsigma_1) + \rho_1) \]

\[ \nabla h^* (\nabla h(\chi_1) + \tau_1) \]

\[ \varsigma_1 = p \p \nabla h^* (\nabla h(\chi_1) + \tau_1) \]

\[ \chi_2 = p \p p \q \nabla h^* (\nabla h(\varsigma_1) + \rho_1) \]

\[ \nabla h^* (\nabla h(\varsigma_1) + \rho_1) \]

\[ \varsigma_0 = p \p q \]
New Result: Euclidean BP is Convergent

- **Euclidean BP:**
  - $f = \| \cdot \|_2^2$
  - $\mathcal{P}, \mathcal{Q}$ arbitrary convex
  - *we have proved converges*!

- Observed to converge near to the desired projection!

- Has strong implications as to the frequent observed good behavior of BP in factor graphs with loops. It is a “curved” version of a provably convergent algorithm. Convergence problems are due to asymmetry of divergence and curvature of sets.
New: Good Behavior of Regular BP for Some Cyclic Factorings

Motivating ideas:

- well known that BP gives correct marginals on acyclic factor graphs (trees and forests)
- how can we translate this to our information geometric framework? (i.e. What region of initializations $\chi_0$ does this correspond to?)
- how can we use the new framework to get a new larger (factor graph girth independent!) set of factorings for which BP gives answers equal or close to true marginals?
  - Popular idea: factor graphs with large girth are “close to” acyclic and (since BP is a message passing algorithm) BP after a finite number of iterations is “close to” the true marginals.
  - Common sense: many other ways for a factoring to be close to acyclic (offending factor in a loop can be weakly dependent on offending variable)
  - information geometry opens up ways to systematize this idea.
Initializations $\chi_0$ From Acyclic Factorings

$$\chi_0 = \prod_{k=1}^{K} g_k(x^k)$$

- Parameterize these by $2^M - 1$ dimensional log coordinates $\theta_k$ of factors $g_k(x^k)$
- Independence of a factor on a variable $\iff \theta_k$ lies in particular vector subspace.
- So a particular acyclic graph is associated with a particular collection of vector subspaces which $\{\theta_k\}$ must live in.
- Set of all log coordinates of acyclic factorings $\mathcal{T}$ is then a union (over all possible acyclic graphs) of vector subspaces!
New: Good Behavior of Regular BP for Some Cyclic Factorings, cont’d

L iterations of BP

acyclic factorings

all distributions

desired projection

product distributions

cyclic factorings

all distributions
New: Good Behavior of Regular BP for Some Cyclic Factorings, cont’d
Conclusions and Future Work

- Information geometric interpretations of BP/turbo decoding are not new. [12, 13, 14, 15, 16]

- Ours is first to make connection with Dykstra’s algorithm

- Formulating belief propagation in the context of Dykstra’s algorithm with Bregman projections allowed us to:
  - prove Euclidean BP always converges.
  - (regular) BP’s occasional convergence problems are a function of the curvature of the sets and the asymmetry of the divergence alone
  - Provide a factor graph girth independent description of factorings for which BP provides answers close to the true marginals after a finite number of iterations.

- A number of more sophisticated convergence and performance conditions are expected to result as well.

- Will use translate condition from new convergence theorem to prove convergence for practical instances of BP decoder (e.g. turbo decoder)
References


