A Completed Information Projection Interpretation of Expectation Propagation

John MacLaren Walsh
Dept. of ECE
Drexel University
Philadelphia, PA 19104
Exponential Family Densities & Rudimentary Information Geometry

- form

\[ p_{\theta}(\theta) = \exp (t(\theta) \cdot \lambda - \psi_t(\lambda)) , \quad \psi_t(\lambda) := \log \left( \int_\Theta \exp (t(\theta) \cdot \lambda - \psi_t(\lambda)) d\theta \right) \]  

(1)

- duality between partition function and neg-entropy

\[ \eta := \int_\Theta t(\theta) \exp (t(\theta) \cdot \lambda - \psi_t(\lambda)) d\theta \]

This map between \( \lambda \) and \( \eta \) is one to one, and may be interpreted to be the gradient of the log partition function

\[ \eta = \nabla_\lambda \psi_t(\lambda) \]

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

\[ h(\eta) := - \inf_{\lambda_1} \{ \psi_t(\lambda_1) - \eta \cdot \lambda_1 \} = \psi_t(\lambda) - \eta \cdot \lambda \]
Expectation Propagation

- joint density

\[ p_{r, \theta}(r, \theta) \propto \prod_{a=1}^{M} f_{a, r}(\theta_a), \quad \theta_a \subseteq \theta \]  

(2)

- approximate

\[ p_{r, \theta}(r, \theta) \approx \prod_{a=1}^{M} g_{a, \lambda_a}(r)(\theta_a) \]  

(3)

- refinement rules

\[ g_{a, \lambda_a} = \text{arg min}_{g_{a, \lambda_a}} \mathcal{D}(v_a \| q) \]

\[ v_a(\theta) := \alpha f_{a, r}(\theta_a) \prod_{c \neq a} g_{c, \lambda_c}(\theta_c), \quad q(\theta) := \beta \prod_{c=1}^{M} g_{c, \lambda_c}(\theta_c) \]  

(4)

\[ \nabla_{\lambda_a} \mathcal{D} = \mathbb{E}_q [t_a(\theta_a)] - \mathbb{E}_{v_a} [t_a(\theta_a)] \]  

(5)
Bregman Divergences

• differentiable convex function is lower bounded by 1st Taylor app. $h$ of Legendre type [1] then

$$h(\chi) \geq h(\varsigma) + \nabla h(\varsigma) \cdot (\chi - \varsigma) \quad (6)$$

with equality if and only if $\chi = \varsigma$.

• Bregman Divergence [1], $B_h$ associated with $h$:

$$B_h(\chi, \varsigma) := h(\chi) - h(\varsigma) - \nabla h(\varsigma) \cdot (\chi - \varsigma)$$

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

$$B_h(\chi, \varsigma) \geq 0 \quad B_h(\chi, \varsigma) = 0 \iff \chi = \varsigma$$

• non-symmetric, triangle inequality in only a subset of cases

• KL Divergence: choose $h$ as negentropy
Method of Alternating Bregman Projections

Find points in 2 convex sets $\mathcal{P}$ and $\mathcal{Q}$ which min. 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* [2, 3, 4] which may be described via the iteration

$$
\chi^{(k)} := \overset{←}{\mathcal{P}} s^{(k)}, \quad s^{(k+1)} := \overset{→}{\mathcal{Q}} \chi^{(k)}
$$
Dykstra’s Algorithm with Cyclic Bregman Projections

\[
\chi^{(k+1)} := \left\lceil \nabla h^* \left( \nabla h(\chi^{(k)}) + \tau^{(k+1-S)} \right) \right\rceil_{C_{kmod}} \quad (7)
\]

\[
\tau^{(k+1)} := \nabla h(\chi^{(k)}) + \tau^{(k+1-S)} - \nabla h(\chi^{(k+1)}) \quad (8)
\]

where we initialize \( \tau^{(-S+1)} , \ldots , \tau^{(0)} = 0 \). This algorithm, under some assumptions, can be shown \([1]\) to solve the best approximation problem, in which one is seeking the point in \( C := \bigcap_{i=0}^{S-1} C_i \) which minimizes the Bregman divergence \( B_h \) in the first argument from the initial point \( \chi^{(0)} \).

*method of cyclic Bregman projections*\([5][6]\) Instead of choosing the convex set to project on cyclicly, one may also choose it randomly \([7, 8]\).
Two Sets Related to EP

• Make as many copies of the space as factors

• One set: densities which are supported on all copies being equal

\[ Q := \{ b \in B | \mathbb{P}_b [x^1 = \cdots = x^M] = 1 \} \]  

(9)

• Another set: product of densities of approximating family form

\[ \mathcal{P} := \{ b | b = \exp (\lambda \cdot \hat{t}(x) - \psi_t(\lambda)), \lambda \in \mathbb{R}^{MV} \} \]  

(10)

• Starting point: one factor per copy
Actual Sets $\mathcal{P}$ & $\mathcal{Q}$ for 2 Bits

Figure 1: The sets $\mathcal{E}_P$ projects between.
EP as a Hybrid Algorithm:

- Desired solution is projection:
  \[
  \mathbf{p}_P \circ \mathbf{p}_Q \left( \frac{\int_{\Theta^M} s(x) \prod_{a=1}^M f_a(x^a) dx}{\int_{\Theta^M} \prod_{a=1}^M f_a(x^a) dx} \right)
  \]  
  \[\quad (11)\]
  \[
  \mathbb{E}_g[t(x^a)] = \mathbb{E}_{\theta|r}[t(\theta)] \quad \forall a \in \{1, \ldots, M\}
  \]

- EP iteratively tries to find it
  \[
  \rho_0, \tau_0 = 0, \quad \chi_0 = \frac{\int_{\Theta^M} s(x) \prod_{a=1}^M f_a(x^a) dx}{\int_{\Theta^M} \prod_{a=1}^M f_a(x^a) dx}, \quad k \in \{0, 1, \ldots, \}
  \]
  \[
  s_k := \mathbf{p}_P \circ \nabla h^* (\nabla h(\chi_k) + \tau_k), \quad \tau_{k+1} := \nabla h(\chi_k) + \tau_k - \nabla h(s_k)
  \]
  \[
  \chi_{k+1} := \mathbf{p}_P \circ \mathbf{p}_Q \circ \nabla h^* (\nabla h(s_k) + \rho_k), \quad \rho_{k+1} := \nabla h(s_k) + \rho_k - \nabla h(\chi_{k+1})
  \]

- Processing for left projection followed by right projection

- Can be viewed as a hybrid between alt. breg. proj. and Dykstra’s w/ cyclic proj.
Figure 2: EP (solid arrows), and the composite projection problem it iteratively solves (dotted arrows), but with $\|\|_2^2$ as the Bregman divergence and different sets.
What does all this mean:

• Alternatively, could say that EP replaces a left proj. w/ a right proj. and log convex for convex from a convergent algorithm (Dykstra’s w/ cyclic proj.)

• Later could be root of occasionally good convergence behavior.

• favorite toy open problem of presenter, yours too now?
Relationship to Prior Work (what was innovated)

• Innovated connection with Dykstra allowed for possible explanation of extrinsic information extraction within context of projection algorithm

\[
g_{a,\lambda_a} = \arg\min_{g_{a,\lambda_a}} \mathcal{D}(v_a \| q)
\]

\[
v_a(\theta) := \alpha f_{a,r}(\theta_a) \prod_{c \neq a} g_{c,\lambda_c}(\theta_c), \quad q(\theta) := \beta \prod_{c=1}^{M} g_{c,\lambda_c}(\theta_c) \tag{12}
\]

• previous expositions had it as an intervening step amidst other projection
So You Don’t Go Home Hungry: Nonlinear Block Gauss Seidel Connection

• write down entire system for EP stationary point

• view Refinement as solution for subset of variables for subset of equations
Convergence Theorem by Applying NLBGS Theory

Relevant references include [9], [10], and [11]. Our next theorem is an application of a theorem from [11]

\[(\lambda_a + \gamma_a) - \Lambda \left( \frac{\int_{\Theta} t_a(\theta_a) \exp \left( \sum_{c=1}^{M} t_c(\theta_c) \cdot \lambda_c \right) d\theta}{\int_{\Theta} \exp \left( \sum_{c=1}^{M} t_c(\theta_c) \cdot \lambda_c \right) d\theta} \right) = 0\]  

\[(\lambda_a + \gamma_a) - \Lambda \left( \frac{\int_{\Theta} u_a(y_a) f_a(y_a) \exp (u_a(y_a) \cdot \gamma_a) dx_a}{\int_{\Theta} f_a(y_a) \exp (u_a(y_a) \cdot \gamma_a) dx_a} \right) = 0\]

Thm. 1 (Convergence of Expectation Propagation Algorithms (Single Parameter Space)): If, when regarded as a function of \([\lambda^T, \gamma^T]^T\), the vector function set equal to zero in the system of equations (13) and (14) is an \(m\)-function, continuous, and surjective (onto) \(\mathbb{R}^{2V}\), the expectation propagation algorithm converges to the unique solution of (13) and (14) and thus the unique interior critical point of the constrained optimization problem.
References


