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

$$\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \exp\left(\mathbf{t}(\boldsymbol{\theta}) \cdot \boldsymbol{\lambda} - \psi_{\mathbf{t}}(\boldsymbol{\lambda})\right), \quad \psi_{\mathbf{t}}(\boldsymbol{\lambda}) := \log\left(\int_{\Theta} \exp\left(\mathbf{t}(\boldsymbol{\theta}) \cdot \boldsymbol{\lambda} - \psi_{\mathbf{t}}(\boldsymbol{\lambda})\right) \mathsf{d}\boldsymbol{\theta}\right) \tag{1}$$

- duality between partition function and neg-entropy

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

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

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

Expectation Propagation

- joint density

$$\mathsf{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r},\boldsymbol{\theta}) \propto \prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathsf{a},\mathbf{r}}(\boldsymbol{\theta}_{\mathsf{a}}), \quad \boldsymbol{\theta}_{\mathsf{a}} \subseteq \boldsymbol{\theta} \tag{2}$$

- approximate

$$\mathsf{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r},\boldsymbol{\theta}) \approx \prod_{\mathsf{a}=1}^{\mathtt{M}} \mathsf{g}_{\mathsf{a},\boldsymbol{\lambda}_{\mathsf{a}}(\mathbf{r})}(\boldsymbol{\theta}_{\mathsf{a}}) \tag{3}$$

- refinement rules

$$\mathsf{g}_{\mathsf{a},\boldsymbol{\lambda}_{\mathsf{a}}} = \arg\min_{\mathsf{g}_{\mathsf{a}},\boldsymbol{\lambda}_{\mathsf{a}}} \mathfrak{D}\left(\mathsf{v}_{\mathsf{a}} \,\|\mathsf{q}\right)$$

$$\mathsf{v}_{\mathsf{a}}(\boldsymbol{\theta}) := \alpha \mathsf{f}_{\mathsf{a},\mathbf{r}}(\boldsymbol{\theta}_{\mathsf{a}}) \prod_{\mathsf{c}\neq\mathsf{a}} \mathsf{g}_{\mathsf{c},\boldsymbol{\lambda}_{\mathsf{c}}}(\boldsymbol{\theta}_{\mathsf{c}}), \;\; \mathsf{q}(\boldsymbol{\theta}) := \beta \prod_{\mathsf{c}=1}^{\mathtt{M}} \mathsf{g}_{\mathsf{c},\boldsymbol{\lambda}_{\mathsf{c}}}(\boldsymbol{\theta}_{\mathsf{c}}) \tag{4}$$

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

Bregman Divergences

- differentiable convex function is lower bounded by 1st Taylor app. h of Legendre type [1] 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 $\boldsymbol{\chi} = \boldsymbol{\varsigma}$.

- *Bregman Divergence* [1], B_{h} associated with h:

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

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

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

- 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

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

Dykstra's Algorithm with Cyclic Bregman Projections

$$\boldsymbol{\chi}^{(\mathtt{k}+1)} := \overleftarrow{\mathbf{p}}_{\mathcal{C}_{\mathtt{k} \bmod \mathtt{S}}} \nabla \mathsf{h}^* \left(\nabla \mathsf{h}(\boldsymbol{\chi}^{(\mathtt{k})}) + \boldsymbol{\tau}^{(\mathtt{k}+1-\mathtt{S})} \right) \tag{7}$$

$$\boldsymbol{\tau}^{(\mathtt{k}+1)} := \nabla \mathsf{h}(\boldsymbol{\chi}^{(\mathtt{k})}) + \boldsymbol{\tau}^{(\mathtt{k}+1-\mathtt{S})} - \nabla \mathsf{h}(\boldsymbol{\chi}^{(\mathtt{k}+1)}) \tag{8}$$

where we initialize $\boldsymbol{\tau}^{(-\mathtt{S}+1)}, \ldots, \boldsymbol{\tau}^{(0)} = \mathbf{0}$. This algorithm, under some assumptions, can be shown [1] 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 $\boldsymbol{\chi}^{(0)}$.

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

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

- Another set: product of densities of approximating family form

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

- Starting point: one factor per copy

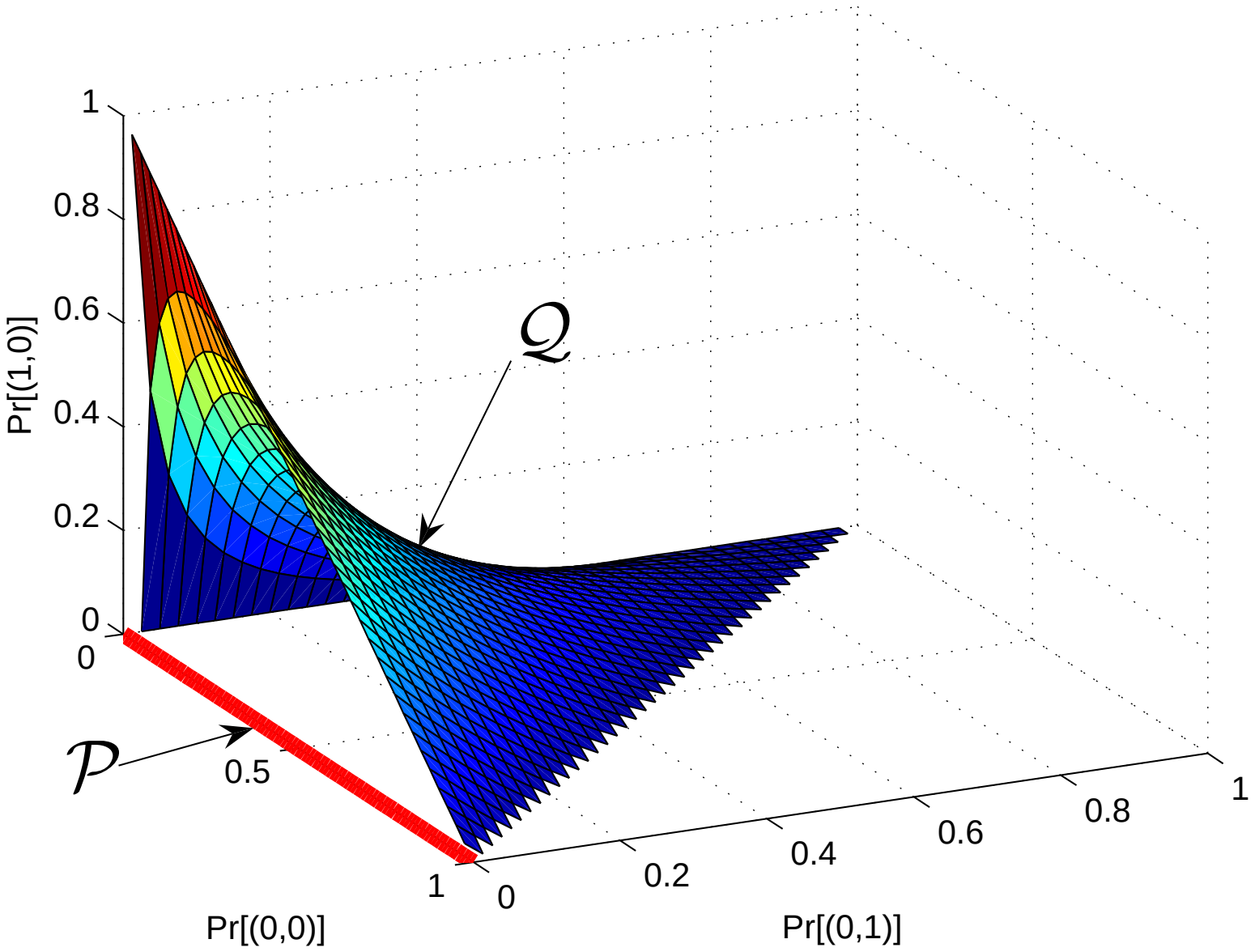

Figure 1: The sets EP projects between.

EP as a Hybrid Algorithm:

- Desired solution is projection:

$$\overrightarrow{\mathbf{p}}_{\mathcal{P}} \circ \overleftarrow{\mathbf{p}}_{\mathcal{Q}} \left(\frac{\int_{\Theta^{\mathtt{M}}} \mathbf{s}(\mathbf{x}) \prod_{\mathtt{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathtt{a}}(\mathbf{x}^{\mathtt{a}}) \mathrm{d}\mathbf{x}}{\int_{\Theta^{\mathtt{M}}} \prod_{\mathtt{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathtt{a}}(\mathbf{x}^{\mathtt{a}}) \mathrm{d}\mathbf{x}} \right) \tag{11}$$

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

- EP iteratively tries to find it

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

- processing for left projection followed by right projection
- Can be viewed as a hybrid between alt. breg. proj. and Dykstra's w/ cyclic proj.

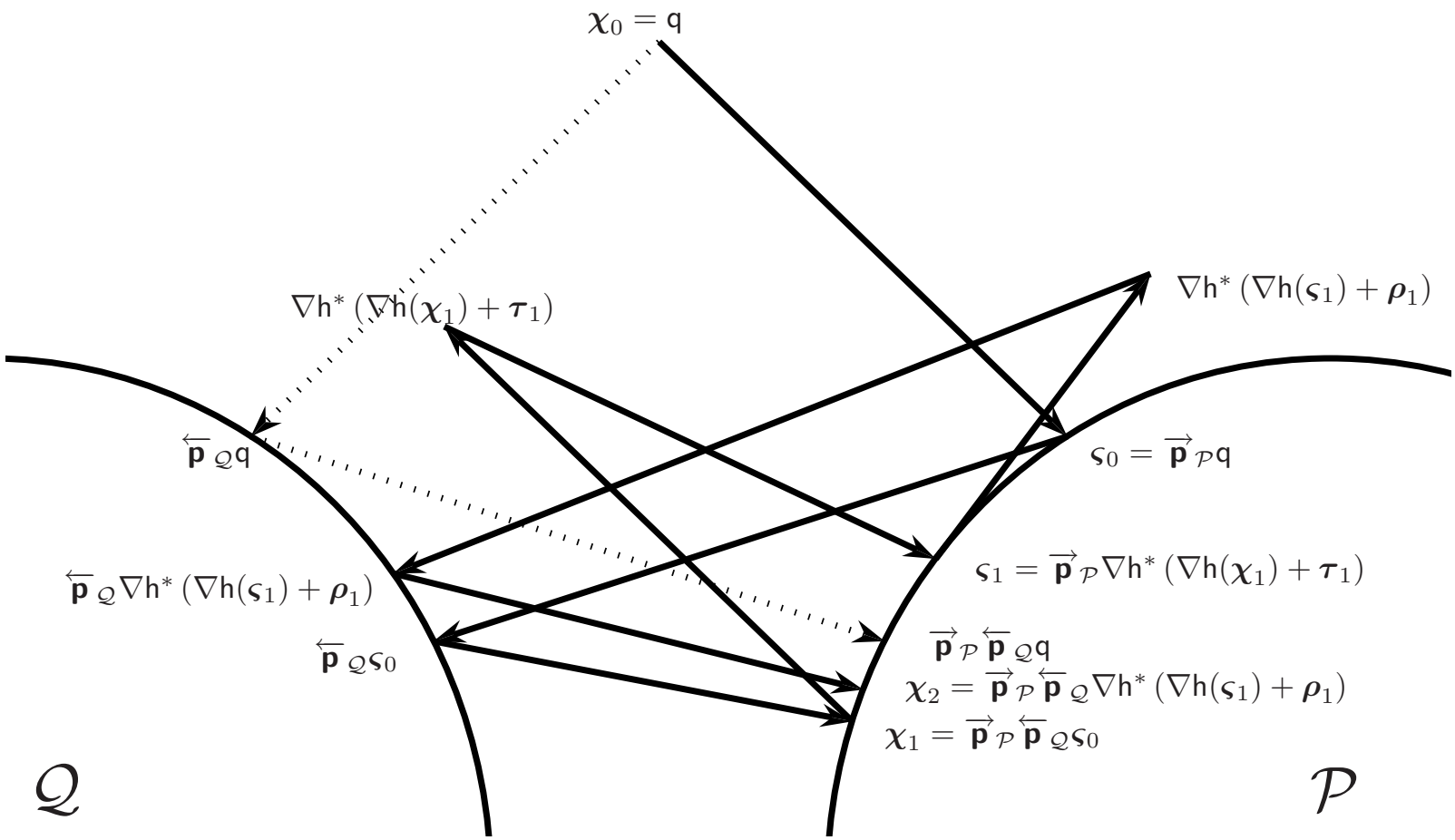

Figure 2: EP (solid arrows), and the composite projection problem it iteratively solves (dotted arrows), but with $\|\|_2^2$ as the Bregman divergence and different sets.

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

$$\mathsf{g}_{\mathsf{a},\boldsymbol{\lambda}_\mathsf{a}} = \arg\min_{\mathsf{g}_{\mathsf{a}},\boldsymbol{\lambda}_\mathsf{a}} \mathfrak{D}\left(\mathsf{v}_\mathsf{a} \,\|\mathsf{q}\right)$$

$$\mathsf{v}_\mathsf{a}(\boldsymbol{\theta}) := \alpha \mathsf{f}_{\mathsf{a},\mathbf{r}}(\boldsymbol{\theta}_\mathsf{a}) \prod_{\mathsf{c}\neq\mathsf{a}} \mathsf{g}_{\mathsf{c},\boldsymbol{\lambda}_\mathsf{c}}(\boldsymbol{\theta}_\mathsf{c}),\ \mathsf{q}(\boldsymbol{\theta}) := \beta \prod_{\mathsf{c}=1}^{\mathtt{M}} \mathsf{g}_{\mathsf{c},\boldsymbol{\lambda}_\mathsf{c}}(\boldsymbol{\theta}_\mathsf{c}) \tag{12}$$

- previous expositions had it as an intervening step admist 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]

$$(\boldsymbol{\lambda}_{\mathtt{a}} + \boldsymbol{\gamma}_{\mathtt{a}}) - \boldsymbol{\Lambda}\left(\frac{\int_{\Theta} \mathbf{t}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}}) \exp\left(\sum_{\mathtt{c}=1}^{\mathtt{M}} \mathbf{t}_{\mathtt{c}}(\boldsymbol{\theta}_{\mathtt{c}}) \cdot \boldsymbol{\lambda}_{\mathtt{c}}\right) \mathrm{d}\boldsymbol{\theta}}{\int_{\Theta} \exp\left(\sum_{\mathtt{c}=1}^{\mathtt{M}} \mathbf{t}_{\mathtt{c}}(\boldsymbol{\theta}_{\mathtt{c}}) \cdot \boldsymbol{\lambda}_{\mathtt{c}}\right) \mathrm{d}\boldsymbol{\theta}}\right) = \mathbf{0} \tag{13}$$

$$(\boldsymbol{\lambda}_{\mathtt{a}} + \boldsymbol{\gamma}_{\mathtt{a}}) - \boldsymbol{\Lambda}\left(\frac{\int_{\Theta} \mathbf{u}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}}) \mathsf{f}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}}) \exp\left(\mathbf{u}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}}) \cdot \boldsymbol{\gamma}_{\mathtt{a}}\right) \mathrm{d}\mathbf{x}_{\mathtt{a}}}{\int_{\Theta} \mathsf{f}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}}) \exp\left(\mathbf{u}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}}) \cdot \boldsymbol{\gamma}_{\mathtt{a}}\right) \mathrm{d}\mathbf{x}_{\mathtt{a}}}\right) = \mathbf{0} \tag{14}$$

Thm. 1 (Convergence of Expectation Propagation Algorithms (Single Parameter Space))**:** If, when regarded as a function of $[\boldsymbol{\lambda}^T, \boldsymbol{\gamma}^T]^T$, the vector function set equal to zero in the system of equations (13) and (14) is an m-function, continuous, and surjective (onto) $\mathbb{R}^{2\mathtt{V}}$, 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

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

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

[3] 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.

[4] 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.

[5] L. M. Bregman, "Proof of the convergence of sheleikhovskii's method for a problem with transportation constraints," *USSR J. Comp. Math. and Math. Phys.*, vol. 7, no. 1, pp. 147–156, 1967.

[6] 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.

[7] H.H. Bauschke and J.M. Borwein, "Legendre functions and the method of random Bregman projections," *J. Conv. Anal.*, , no. 4(1), pp. 27–67, 1997.

[8] H.H. Bauschke and D. Noll, "The method of forward projections," *J. Nonlin. and Conv. Anal.*, vol. 3, no. 2, pp. 191–205, 2002.

[9] W. C. Rheinboldt, "On m-functions and their application to nonlinear gauss-seidel iterations and network flows," *Journal Mathematical Analysis and Applications*, pp. 274–307, 1971.

[10] J. J. Moré, "Nonlinear generalizations of matrix diagonal dominance with application to gauss-seidel iterations," *SIAM Journal on Numerical Analysis*, vol. 9, pp. 357–378, June 1972.

[11] W. C. Rheinboldt, "On classes of n-dimensional nonlinear mappings generalizing several types of matrices," in *Proc. Symposium on the Numerical Solution of Partial Differential Equations. II.* 1970, pp. 501–546, Academic Press.