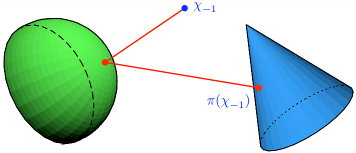


Information Projection Algorithms and Belief Propagation

Phil Regalia

Department of Electrical Engineering and Computer Science
Catholic University of America
Washington, DC 20064

with J. M. Walsh, Drexel University

P. Dewilde Workshop, Wassenaar, June 2008

Outline

Belief Propagation $\Longleftrightarrow$ Iterative Convex Projection algorithms

Belief Propagation (Pearl, 1986) has met with success in:

- Error correction decoding (low density parity-check codes, turbo codes, . . .);
- Network diagnostics and link monitoring;
- Sensor self-localization;
- Distributed estimation in sensor networks;
- Lossy source quantization;
- Multi-user communications;
- Et cetera.

Iterative Convex Projection algorithms have convergence proofs, unlike BP for which proofs assume either:

- Tree or forest dependency graph;
- Arbitrarily large factor graph girth.

Belief Propagation $\Longleftrightarrow$ Iterative Convex Projection algorithms

Belief Propagation (Pearl, 1986) has met with success in:

- Error correction decoding (low density parity-check codes, turbo codes, ...);
- Network diagnostics and link monitoring;
- Sensor self-localization;
- Distributed estimation in sensor networks;
- Lossy source quantization;
- Multi-user communications;
- Et cetera.

Iterative Convex Projection algorithms have convergence proofs, unlike BP for which proofs assume either:

- Tree or forest dependency graph;
- Arbitrarily large factor graph girth.

Belief Propagation $\Longleftrightarrow$ Iterative Convex Projection algorithms

Belief Propagation (Pearl, 1986) has met with success in:

- Error correction decoding (low density parity-check codes, turbo codes, . . .);
- Network diagnostics and link monitoring;
- Sensor self-localization;
- Distributed estimation in sensor networks;
- Lossy source quantization;
- Multi-user communications;
- Et cetera.

Iterative Convex Projection algorithms have convergence proofs, unlike BP for which proofs assume either:

- Tree or forest dependency graph;
- Arbitrarily large factor graph girth.

Basic Query

Can iterative convex projection algorithms lend insight into belief propagation?

A priori yes, and the role of information projections and information geometry in BP is nothing new:

- Moher and Gulliver (Trans. IT-98) rephrased iterative decoding as information projections of Csiszár;
- Grant (ISIT-99) examined turbo decoding via information projections;
- Richardson (Trans. IT-03) viewed nonlinear dynamics of iterative decoding via (tacitly) information geometry;
- Ikeda, Tanaka and Amari (Trans. IT-04) rephrased "all of the above" in terms of information geometry.

Key obstacle: "Extrinsic information extraction" goes against projections on invariant sets.

Basic Query

Can iterative convex projection algorithms lend insight into belief propagation?

A priori yes, and the role of information projections and information geometry in BP is nothing new:

- Moher and Gulliver (Trans. IT-98) rephrased iterative decoding as information projections of Csiszár;
- Grant (ISIT-99) examined turbo decoding via information projections;
- Richardson (Trans. IT-03) viewed nonlinear dynamics of iterative decoding via (tacitly) information geometry;
- Ikeda, Tanaka and Amari (Trans. IT-04) rephrased "all of the above" in terms of information geometry.

Key obstacle: "Extrinsic information extraction" goes against projections on invariant sets.

Basic Query

Can iterative convex projection algorithms lend insight into belief propagation?

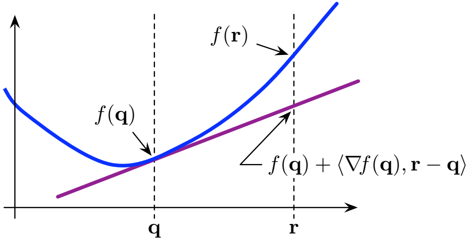
A priori yes, and the role of information projections and information geometry in BP is nothing new:

- Moher and Gulliver (Trans. IT-98) rephrased iterative decoding as information projections of Csiszár;
- Grant (ISIT-99) examined turbo decoding via information projections;
- Richardson (Trans. IT-03) viewed nonlinear dynamics of iterative decoding via (tacitly) information geometry;
- Ikeda, Tanaka and Amari (Trans. IT-04) rephrased "all of the above" in terms of information geometry.

Key obstacle: "Extrinsic information extraction" goes against projections on invariant sets.

Bregman Distance

Graph of a convex function $f(\cdot)$ is lower bounded by any tangent hyperplane:



This induces the gradient inequality:

$$f(\mathbf{r}) \geq f(\mathbf{q}) + \langle \nabla f(\mathbf{q}), \mathbf{r} - \mathbf{q} \rangle$$

whose discrepancy in turn induces the *Bregman distance*

$$D_f(\mathbf{r}, \mathbf{q}) \triangleq f(\mathbf{r}) - f(\mathbf{q}) - \langle \nabla f(\mathbf{q}), \mathbf{r} - \mathbf{q} \rangle \geq 0$$

Common examples

Probability mass functions defined on M outcomes:

$$q_i = \Pr(x = x_i), \qquad i = 0, 1, 2, \ldots, M{-}1.$$

Introduce convex domain

$$\mathcal{D} = \left\{ q_i : \begin{array}{c} q_1 \geq 0 \\ \vdots \\ q_{M-1} \geq 0 \end{array}, \quad q_1 + q_2 + \cdots + q_{M-1} \leq 1 \right\}$$

Setting $q_0 = 1 - \sum_{i=1}^{M-1} q_i$, the negative Shannon entropy

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

is convex over $\mathcal{D}$.

The induced Bregman distance becomes

$$D_f(\mathbf{r}, \mathbf{q}) = \sum_{i=0}^{M-1} r_i \log \frac{r_i}{q_i}$$

and is recognized as the Kullback-Leibler divergence.

Closely related is the Fenchel conjugate function

$$\begin{aligned} f^*(\boldsymbol{\theta}) &\triangleq \sup_{\mathbf{q}} \Big(\langle \mathbf{q}, \boldsymbol{\theta} \rangle - f(\mathbf{q}) \Big) \\ &= \log \left(\sum_{i=0}^{M-1} \exp(\theta_i) \right), \qquad \text{with } \theta_i = \log \frac{q_i}{q_0} \end{aligned}$$

This is convex over a domain $\mathcal{D}^* \in \mathbb{R}^M$ of vectors having first component zero. ("Log probability ratios")

CUA

The induced Bregman distance becomes

$$D_f(\mathbf{r}, \mathbf{q}) = \sum_{i=0}^{M-1} r_i \log \frac{r_i}{q_i}$$

and is recognized as the Kullback-Leibler divergence.

Closely related is the Fenchel conjugate function

$$\begin{aligned} f^*(\boldsymbol{\theta}) &\triangleq \sup_{\mathbf{q}} \left(\langle \mathbf{q}, \boldsymbol{\theta} \rangle - f(\mathbf{q}) \right) \\ &= \log \left(\sum_{i=0}^{M-1} \exp(\theta_i) \right), \qquad \text{with } \theta_i = \log \frac{q_i}{q_0} \end{aligned}$$

This is convex over a domain $\mathcal{D}^* \in \mathbb{R}^M$ of vectors having first component zero. ("Log probability ratios")

Induced Bregman distance is now

$$D_{f^*}(\boldsymbol{\rho}, \boldsymbol{\theta}) = \log \frac{\sum_{i=0}^{M-1} \exp(\rho_i)}{\sum_{i=0}^{M-1} \exp(\theta_i)} - \sum_{i=0}^{M-1} \frac{\exp(\theta_i)\,(\rho_i - \theta_i)}{\sum_{j=0}^{M-1} \exp(\theta_j)}$$

By associating

$$q_i = \frac{\exp(\theta_i)}{\sum_{j=0}^{M-1} \exp(\theta_j)} \qquad r_i = \frac{\exp(\rho_i)}{\sum_{j=0}^{M-1} \exp(\rho_j)}$$

This assumes the more familiar form

$$D_{f^*}(\boldsymbol{\rho}, \boldsymbol{\theta}) = \sum_{i=0}^{M-1} q_i \log \frac{q_i}{r_i}$$

Observe that

$$
\begin{aligned}
[\nabla f(\mathbf{q})]_i &= \frac{df(\mathbf{q})}{dq_i} = \log \frac{q_i}{q_0} = \theta_i \\
[\nabla f^*(\boldsymbol{\theta})]_i &= \frac{df^*(\boldsymbol{\theta})}{d\theta_i} = \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)} = q_i
\end{aligned}
$$

giving inverse maps, as $f(\mathbf{q})$ and $f^*(\boldsymbol{\theta})$ are Legendre transforms of each other.

In general, for convex f and its conjugate f^* in the Legendre class, their induced Bregman distances satisfy

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

Consider finally the energy function:

$$f(\mathbf{q}) = \frac{1}{2}\sum_{i=0}^{M-1} {q_i}^2, \qquad \mathbf{q} \in \mathbb{R}^M$$

the Bregman distance becomes

$$D_f(\mathbf{r}, \mathbf{q}) = \frac{1}{2}\sum_{i=0}^{M-1} (r_i - q_i)^2.$$

As $\nabla f(\mathbf{q}) = \mathbf{q}$, and $f^* = f$, this gives a symmetric Bregman distance:

$$D_f(\mathbf{r}, \mathbf{q}) = D_f(\mathbf{q}, \mathbf{r})$$

Bregman Projections

If $f(\mathbf{q})$ is a convex function over a domain $\mathcal{D}$, and $\mathcal{C}$ is a convex subset of $\mathcal{D}$, the Bregman projection of $\mathbf{q}$ onto $\mathcal{C}$ is

$$\pi_{\mathcal{C}}(\mathbf{q}) = \arg\min_{\mathbf{r}\in\mathcal{C}} D_f(\mathbf{r}, \mathbf{q}).$$

and is characterized by the inequality

$$D_f(\mathbf{r}, \mathbf{q}) \geq D_f\Big(\mathbf{r}, \pi_{\mathcal{C}}(\mathbf{q})\Big) + D_f\Big(\pi_{\mathcal{C}}(\mathbf{q}), \mathbf{q}\Big), \quad \text{for all } \mathbf{r} \in \mathcal{C},$$

or, equivalently,

$$\Big\langle \nabla f(\mathbf{q}) - \nabla f\big(\pi_{\mathcal{C}}(\mathbf{q})\big), \mathbf{r} - \pi_{\mathcal{C}}(\mathbf{q}) \Big\rangle \leq 0, \qquad \text{for all } \mathbf{r} \in \mathcal{C}.$$

Dykstra's Cyclic Projection Algorithm

Seek minimization

$$\pi_{\mathcal{C}}(\mathbf{q}) = \arg\min_{\mathbf{r}\in\mathcal{C}} D_f(\mathbf{r},\mathbf{q})$$

where $\mathcal{C}$ is the intersection of convex sets: $\mathcal{C} = \bigcap_{n=1}^{N} \mathcal{C}_n$

First stab: Sometimes convergent algorithm,

$$\begin{aligned} \mathbf{r}_n &= \pi_{\mathcal{C}_n}(\mathbf{r}_{n-1}) \\ &= \pi_{\mathcal{C}_n}\left(\nabla f^*(\nabla f(\mathbf{r}_{n-1}))\right) \end{aligned}$$

using $\mathcal{C}_{n+N} = \mathcal{C}_n$ and initialization

$$\mathbf{r}_0 = \mathbf{q},$$

Dykstra's Cyclic Projection Algorithm

Seek minimization

$$\pi_{\mathcal{C}}(\mathbf{q}) = \arg\min_{\mathbf{r}\in\mathcal{C}} D_f(\mathbf{r}, \mathbf{q})$$

where $\mathcal{C}$ is the intersection of convex sets: $\mathcal{C} = \bigcap_{n=1}^{N} \mathcal{C}_n$

Improved algorithm, $\mathbf{r}_n \xrightarrow{n\to\infty} \pi_{\mathcal{C}}(\mathbf{q})$:

$$\begin{aligned} \mathbf{r}_n &= \pi_{\mathcal{C}_n}\left(\nabla f^*(\nabla f(\mathbf{r}_{n-1}) + \mathbf{s}_{n-N})\right) \\ \mathbf{s}_n &= \nabla f(\mathbf{r}_{n-1}) + \mathbf{s}_{n-N} - \nabla f(\mathbf{r}_n) \end{aligned}$$

using $\mathcal{C}_{n+N} = \mathcal{C}_n$ and initialization

$$\mathbf{r}_0 = \mathbf{q}, \qquad \mathbf{s}_{-(N-1)} = \cdots = \mathbf{s}_{-1} = \mathbf{s}_0 = \mathbf{0}.$$

CUA

Minimum Distance Algorithm

Find closest members:

$$D_f(\mathbf{r}_*, \mathbf{q}_*) = \inf_{\mathbf{r}\in\mathcal{C}_1,\, \mathbf{q}\in\mathcal{C}_2} D_f(\mathbf{r}, \mathbf{q}), \qquad \text{where } \mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset.$$

Cyclic projection algorithm becomes

$$\begin{aligned}
\mathbf{r}_n &= \pi_{\mathcal{C}_1}\Big(\nabla f^*\big(\nabla f(\mathbf{q}_{n-1}) + \mathbf{v}_{n-1}\big)\Big) \\
\mathbf{v}_n &= \nabla f(\mathbf{q}_{n-1}) + \mathbf{v}_{n-1} - \nabla f(\mathbf{r}_n) \\
\mathbf{q}_n &= \pi_{\mathcal{C}_2}\Big(\nabla f\big(\nabla f^*(\mathbf{r}_n) + \mathbf{w}_{n-1}\big)\Big) \\
\mathbf{w}_n &= \nabla f^*(\mathbf{r}_n) + \mathbf{w}_{n-1} - \nabla f^*(\mathbf{q}_n)
\end{aligned}$$

Convergent in the Euclidean case $D_f(\mathbf{r}, \mathbf{q}) = \frac{1}{2}\sum_k (r_k - q_k)^2$.

Belief Propagation

Iterative (sometimes "fast") algorithm to calculate marginal probability functions.
Given binary vector $\mathbf{x} = [x_1, \ldots, x_M] \in \{0, 1\}^M$, and probability function $G(\mathbf{x})$, seek marginals

$$f_k(x_k) = \sum_{x_1=0}^{1} \cdots \sum_{x_{k-1}=0}^{1} \sum_{x_{k+1}=0}^{1} \cdots \sum_{x_M=0}^{1} G(\mathbf{x})$$

Calculation is "hard" since there are 2^M evaluations for $\mathbf{x}$.

BP is applicable when $G(\mathbf{x})$ splits into simpler factors:

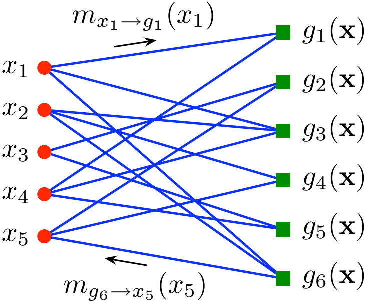
$$G(\mathbf{x}) = \prod_{k=1}^{K} g_k(\mathbf{x})$$

Usually, each factor g_k depends only on a subset of variables in $\mathbf{x}$.

Message passing algorithm

$$m_{x_i \to g_k}(x_i) = \text{variable } x_i \text{ to factor } g_k = \beta_k \prod_{\ell \neq k} m_{g_\ell \to x_i}(x_i)$$

$$m_{g_k \to x_i}(x_i) = \text{factor } g_k \text{ to variable } x_i = \alpha_i \sum_{x_n : n \neq i} g_k(\mathbf{x}) \prod_{\ell \neq i} m_{x_\ell \to g_k}(x_\ell)$$



Outgoing message on an edge depends only on incoming messages from *other* edges at that node (“extrinsic” information).

If convergent, the beliefs

$$b_i(x_i) = \beta_i \prod_{\ell} m_{g_\ell \to x_i}(x_i)$$

are thresholded: $b_i(0) \gtrless b_i(1)$. Ideally, beliefs would be marginals.

For projection interpretation, let $\mathbf{x}_1, \ldots, \mathbf{x}_K$ be K copies of $\mathbf{x}$, and consider extended likelihood function

$$G(\mathbf{x}_1, \ldots, \mathbf{x}_K) = \prod_{k=1}^{K} g_k(\mathbf{x}_k).$$

Original function is obtained with constraint $\mathbf{x}^1 = \cdots = \mathbf{x}^K = \mathbf{x}$.
Two convex sets:

- Product distributions of 2^{MK} variables:

$$\mathcal{P} = \left\{ \boldsymbol{\theta} : \nabla_{f*}(\boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{m=1}^{M} q_{k,m}(x_m^k) \right\}$$

- Constraint (or "sparse") distributions

$$\mathcal{Q} = \left\{ \mathbf{r} : r(\mathbf{x}^1, \ldots, \mathbf{x}^K) = 0 \text{ if } \mathbf{x}^i \neq \mathbf{x}^j \text{ for any } i \neq j \right\}$$

Information geometric view:

- *At factor nodes*: Given pmf $\mathbf{q}$, if $\mathbf{r} \in \mathcal{P}$ is a product distribution built from the marginals of $\mathbf{q}$, then

$$D_f(\mathbf{q}, \mathbf{s}) = D_f(\mathbf{q}, \mathbf{r}) + \underbrace{D_f(\mathbf{r}, \mathbf{s})}_{\geq 0}, \qquad \text{for all } \mathbf{s} \in \mathcal{P}$$

so that $\mathbf{r}$ is the Bregman projection of $\mathbf{q}$ onto $\mathcal{P}$.

- *At variable nodes*: Given pmf $\mathbf{q}$, if $\mathcal{Q}$ denotes a set of "sparse" distributions, then

$$b_i = \begin{cases} \beta\, q_i, & i \in \text{index set for } \mathcal{Q}; \\ 0, & \text{otherwise;} \end{cases}$$

satisfies

$$D_f(\mathbf{s}, \mathbf{q}) = \underbrace{D_f(\mathbf{s}, \mathbf{b})}_{\geq 0} + D_f(\mathbf{b}, \mathbf{q}), \qquad \text{for all } \mathbf{s} \in \mathcal{Q}.$$

so that $\mathbf{b}$ (containing beliefs) is Bregman projection onto $\mathcal{Q}$.

Information geometric view:

- *At factor nodes*: Given pmf $\mathbf{q}$, if $\mathbf{r} \in \mathcal{P}$ is a product distribution built from the marginals of $\mathbf{q}$, then

$$D_f(\mathbf{q}, \mathbf{s}) = D_f(\mathbf{q}, \mathbf{r}) + \underbrace{D_f(\mathbf{r}, \mathbf{s})}_{\geq 0}, \qquad \text{for all } \mathbf{s} \in \mathcal{P}$$

  so that $\mathbf{r}$ is the Bregman projection of $\mathbf{q}$ onto $\mathcal{P}$.

- *At variable nodes*: Given pmf $\mathbf{q}$, if $\mathcal{Q}$ denotes a set of "sparse" distributions, then

$$b_i = \begin{cases} \beta\, q_i, & i \in \text{index set for } \mathcal{Q}; \\ 0, & \text{otherwise}; \end{cases}$$

  satisfies

$$D_f(\mathbf{s}, \mathbf{q}) = \underbrace{D_f(\mathbf{s}, \mathbf{b})}_{\geq 0} + D_f(\mathbf{b}, \mathbf{q}), \qquad \text{for all } \mathbf{s} \in \mathcal{Q}.$$

  so that $\mathbf{b}$ (containing beliefs) is Bregman projection onto $\mathcal{Q}$.

Calculation of marginal pmfs

"Ideal" solution is

$$\begin{aligned} \mathbf{p} &= \pi_{\mathcal{Q}}\big(\nabla f^*(\chi_{-1})\big) \qquad \text{(constrain to sparse distributions)} \\ \chi_0 &= \pi_{\mathcal{P}}\big(\nabla f(\mathbf{p})\big) \qquad \text{(calculate marginal distributions)} \end{aligned}$$

with "initialization"

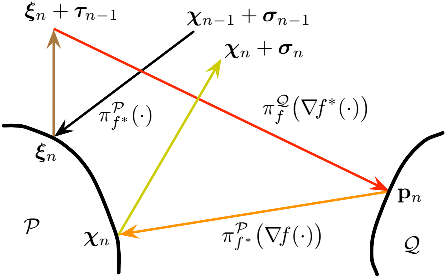
$$\chi_{-1} = \nabla f\left(\prod_k g_k(\mathbf{x}_k)\right)$$

Belief propagation is the cyclic projection algorithm

$$
\begin{aligned}
\boldsymbol{\xi}_n &= \pi_{\mathcal{P}}(\boldsymbol{\chi}_{n-1} + \boldsymbol{\sigma}_{n-1}) \\
\boldsymbol{\sigma}_n &= \boldsymbol{\chi}_{n-1} + \boldsymbol{\sigma}_{n-1} - \boldsymbol{\xi}_n \\
\mathbf{p}_n &= \pi_{\mathcal{Q}}\left(\nabla f^*(\boldsymbol{\xi}_n + \boldsymbol{\tau}_{n-1})\right) \\
\boldsymbol{\chi}_n &= \pi_{\mathcal{P}}\left(\nabla f(\mathbf{p}_n)\right) \\
\boldsymbol{\tau}_n &= \boldsymbol{\xi}_n + \boldsymbol{\tau}_{n-1} - \boldsymbol{\chi}_n
\end{aligned}
$$

with initialization

$$
\begin{aligned}
\boldsymbol{\chi}_{-1} &= \nabla f\left(\prod_k g_k(\mathbf{x}_k)\right) \\
\boldsymbol{\sigma}_{-1} &= \boldsymbol{\tau}_{-1} = \mathbf{0}
\end{aligned}
$$



Special Case: Replace negative Shannon entropy by energy function, to get "Euclidean belief propagation":

$$\begin{aligned}
\boldsymbol{\xi}_n &= \pi_{\mathcal{P}}(\boldsymbol{\chi}_{n-1} + \boldsymbol{\sigma}_{n-1}) \\
\boldsymbol{\sigma}_n &= \boldsymbol{\chi}_{n-1} + \boldsymbol{\sigma}_{n-1} - \boldsymbol{\xi}_n \\
\mathbf{p}_n &= \pi_{\mathcal{Q}}(\boldsymbol{\xi}_n + \boldsymbol{\tau}_{n-1}) \\
\boldsymbol{\chi}_n &= \pi_{\mathcal{P}}(\mathbf{p}_n) \\
\boldsymbol{\tau}_n &= \boldsymbol{\xi}_n + \boldsymbol{\tau}_{n-1} - \boldsymbol{\chi}_n
\end{aligned}$$

Given arbitary convex sets $\mathcal{P}$ and $\mathcal{Q}$, can show that $\{\boldsymbol{\chi}_n\}$ converges for any initial condition $\boldsymbol{\chi}_{-1}$.

Conventional belief propagation, by contrast, is convergent when either of the following apply:

- Factor graph is a tree/forest (or nearly so: large girth);
- Initialization χ_{-1} is a product distribution (or nearly so: $\pi_{\mathcal{P}}(\chi_{-1}) \approx \chi_{-1}$).

CUA

Special Case: Replace negative Shannon entropy by energy function, to get "Euclidean belief propagation":

$$\begin{aligned}
\boldsymbol{\xi}_n &= \pi_{\mathcal{P}}(\boldsymbol{\chi}_{n-1} + \boldsymbol{\sigma}_{n-1}) \\
\boldsymbol{\sigma}_n &= \boldsymbol{\chi}_{n-1} + \boldsymbol{\sigma}_{n-1} - \boldsymbol{\xi}_n \\
\mathbf{p}_n &= \pi_{\mathcal{Q}}(\boldsymbol{\xi}_n + \boldsymbol{\tau}_{n-1}) \\
\boldsymbol{\chi}_n &= \pi_{\mathcal{P}}(\mathbf{p}_n) \\
\boldsymbol{\tau}_n &= \boldsymbol{\xi}_n + \boldsymbol{\tau}_{n-1} - \boldsymbol{\chi}_n
\end{aligned}$$

Given arbitary convex sets $\mathcal{P}$ and $\mathcal{Q}$, can show that $\{\boldsymbol{\chi}_n\}$ converges for any initial condition $\boldsymbol{\chi}_{-1}$.

Conventional belief propagation, by contrast, is convergent when either of the following apply:

- Factor graph is a tree/forest (or nearly so: large girth);
- Initialization $\boldsymbol{\chi}_{-1}$ is a product distribution (or nearly so: $\pi_{\mathcal{P}}(\boldsymbol{\chi}_{-1}) \approx \boldsymbol{\chi}_{-1}$).

CUA

Further perspectives:

- Rephrasing belief propagation in terms of cyclic convex projection algorithms suggests that convergence studies of the latter might extend to the former.
- “Euclidean” belief propagation is always convergent, although conventional (information projection-based) belief propagation can diverge is some settings.
- Continuity of projectors can be invoked to extend cases where belief propagation is proved to converge to “good” solutions.
- Can some modification to belief propagation give a more faithful transcription of Dykstra’s algorithm?


MERRIE MELODIES
"That's all Folks!"
A WARNER BROS. CARTOON
A VITAPHONE RELEASE