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

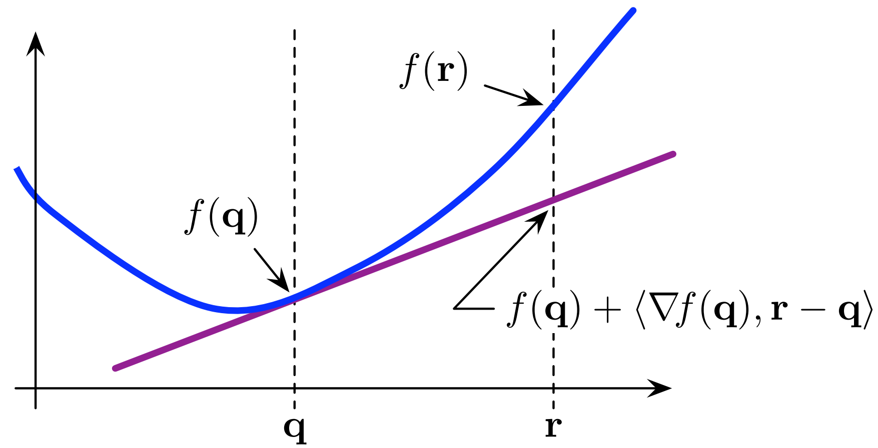
1. Some Convex Programming
   (a) Bregman Divergence
   (b) Bregman Projections & Examples
   (c) Bregman Projections Algorithms: Alternating Bregman Projections & Dykstra's Algorithm with Cyclic Bregman Projections
2. Belief Propagation
   (a) Definition
   (b) Applications (efficient decoders for near 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

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

for all $\mathbf{r} \in \mathcal{D}$



- f *strictly convex*, then strict inequality unless $\mathbf{r} = \mathbf{q}$. Can then use this to construct the **Bregman divergence**

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

  which vanishes $\Leftrightarrow \mathbf{r} = \mathbf{q}$.

- Need not be symmetric, i.e. in general $D_f(\mathbf{q}, \mathbf{r}) \neq D_f(\mathbf{r}, \mathbf{q})$.
- Need not satisfy triangle inequality. (only happens in special cases)

Bregman Divergence, cont'd

- Given convex $f(\mathbf{q})$, have **convex conjugate**

$$f^*(\boldsymbol{\theta}) = \sup_{\mathbf{q}} \Big(\langle \mathbf{q}, \boldsymbol{\theta} \rangle - f(\mathbf{q}) \Big).$$

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

$$\nabla f^*(\nabla f(\mathbf{q})) = \mathbf{q},\ \nabla f(\nabla f^*(\boldsymbol{\theta})) = \boldsymbol{\theta}$$

- Conjugation switches arguments

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

- **Example: Euclidean** $f(\mathbf{q}) = \frac{1}{2}\|\mathbf{q}\|_2^2 = f^*(\mathbf{q})$, $D_f(\mathbf{r}, \mathbf{q}) = \frac{1}{2}\|\mathbf{r} - \mathbf{q}\|_2^2$

Bregman Divergence, cont'd

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

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

the partition function

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

with domain $\mathcal{D}^* = \mathbb{R}_e^N$

Bregman Projections

Because of asymmetry given a Bregman divergence have two notions of projections [2, 3] **Left Projection**: $\mathcal{C} \subset \mathcal{D}$, convex

$$\overleftarrow{\mathbf{P}}_{\mathcal{C}}\mathbf{q} := \arg\min_{\mathbf{r}\in\mathcal{C}} D_f(\mathbf{r}, \mathbf{q})$$

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

$$\overrightarrow{\mathbf{P}}_{\mathcal{P}}\mathbf{q} := \arg\min_{\mathbf{r}\in\mathcal{P}} D_f(\mathbf{q}, \mathbf{r})$$

Right projection with D_f can also be considered as left projection with D_{f^*}.

Bregman Projections Algorithms

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

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

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

- **Dykstra's Algorithm with Cyclic Bregman Projections** [7, 8, 9]

$$\begin{aligned} \boldsymbol{\chi}^{(\mathtt{k}+1)} &:= \overleftarrow{\mathbf{P}}_{\mathcal{C}_{\mathtt{k} \bmod \mathtt{S}}} \nabla f^* \left(\nabla f(\boldsymbol{\chi}^{(\mathtt{k})}) + \boldsymbol{\tau}^{(\mathtt{k}+1-\mathtt{S})} \right) \\ \boldsymbol{\tau}^{(\mathtt{k}+1)} &:= \nabla f(\boldsymbol{\chi}^{(\mathtt{k})}) + \boldsymbol{\tau}^{(\mathtt{k}+1-\mathtt{S})} - \nabla f(\boldsymbol{\chi}^{(\mathtt{k}+1)}) \end{aligned}$$

  with $\boldsymbol{\tau}^{(-\mathtt{S}+1)}, \ldots, \boldsymbol{\tau}^{(0)} = \mathbf{0}$. Under some (benign) assumptions, [7] solves *best approximation problem*,

$$\arg \min_{\boldsymbol{\chi} \in \bigcap_{\mathtt{i}=0}^{\mathtt{S}-1} \mathcal{C}_{\mathtt{i}}} D_f(\boldsymbol{\chi}, \boldsymbol{\chi}^0)$$

- 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 $\mathbf{x}$.

$$m^{(j)}_{x_i \to g_k}(x_i) = \beta_k \prod_{\substack{\ell=1 \\ \ell \neq k}}^{K} m^{(j)}_{g_\ell \to x_i}(x_i)$$

$k = 1, 2, \ldots, K;$

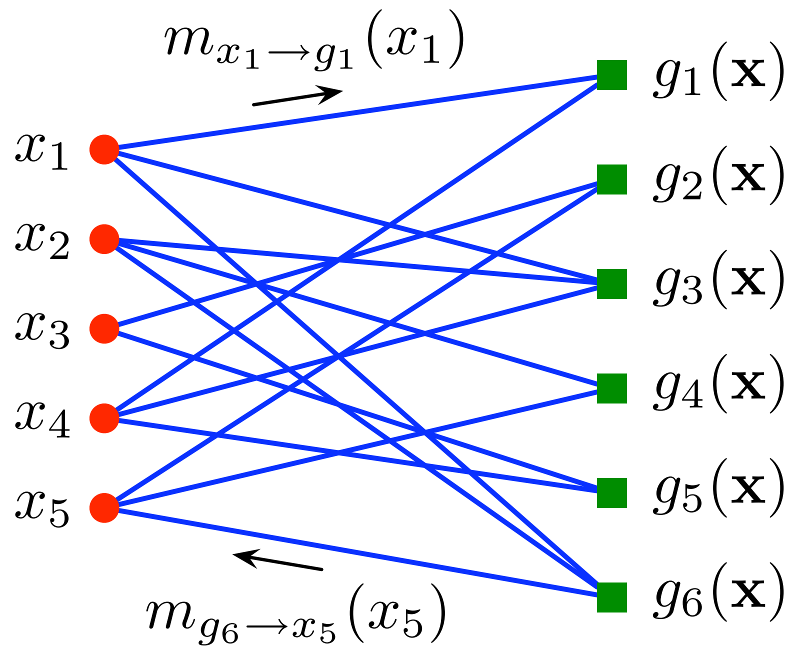


Figure 1: A factor graph.

$$m^{(j)}_{g_k \to x_i}(x_i) = \alpha_i \sum_{x_\ell : \ell \neq i} g_k(\mathbf{x}) \prod_{\substack{n=1 \\ n \neq i}}^{M} m^{(j-1)}_{x_n \to g_k}(x_n), \qquad i = 1, 2, \ldots, M;$$

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"

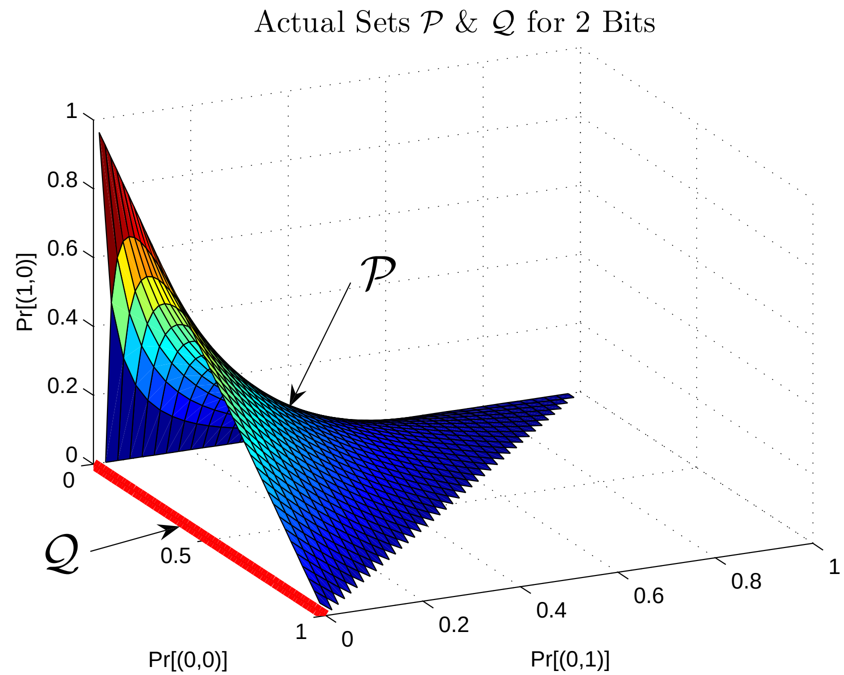
Belief Propagation as a Hybrid Projections Algorithm

- Make K indep. copies of $\mathbf{x}$, forming $\bar{\mathbf{x}} := \mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^K$.
- $\mathbf{q} \in \mathcal{D}$, $\boldsymbol{\theta} \in \mathcal{D}^*$ parameterize the set of PMFs (subset of $N = 2^{KM} - 1$ dimensional space.)
- f the negative Shannon entropy.
- $\mathcal{Q} \subset \mathcal{D}$ the set of $\mathbf{q}$ such that *all K copies are equal*

$$\mathcal{Q} := \left\{\mathbf{q} \middle| \mathbb{P}_{\mathbf{q}}[\mathbf{x}^1 = \cdots = \mathbf{x}^K] = 1\right\}$$

- $\mathcal{P}^* \subset \mathcal{D}^*$ the set of log coordinates dual to the set of product distributions over all bits

$$\mathcal{P} := \left\{ q(\mathbf{x}^1, \ldots, \mathbf{x}^K) = \prod_{k=1}^{K} \prod_{m=1}^{M} q(x_m^k) \right\}$$



Initial point is based on the factoring:

$$\boldsymbol{\chi}_0 = \prod_{k=1}^{K} g_k(\mathbf{x}^k)$$

Belief Propagation as a Hybrid Projections Algorithm

- Desired solution (exact marginals) is two step projection

$$\overrightarrow{\mathbf{P}}_{\mathcal{P}} \circ \overleftarrow{\mathbf{P}}_{\mathcal{Q}} \boldsymbol{\chi}_0$$

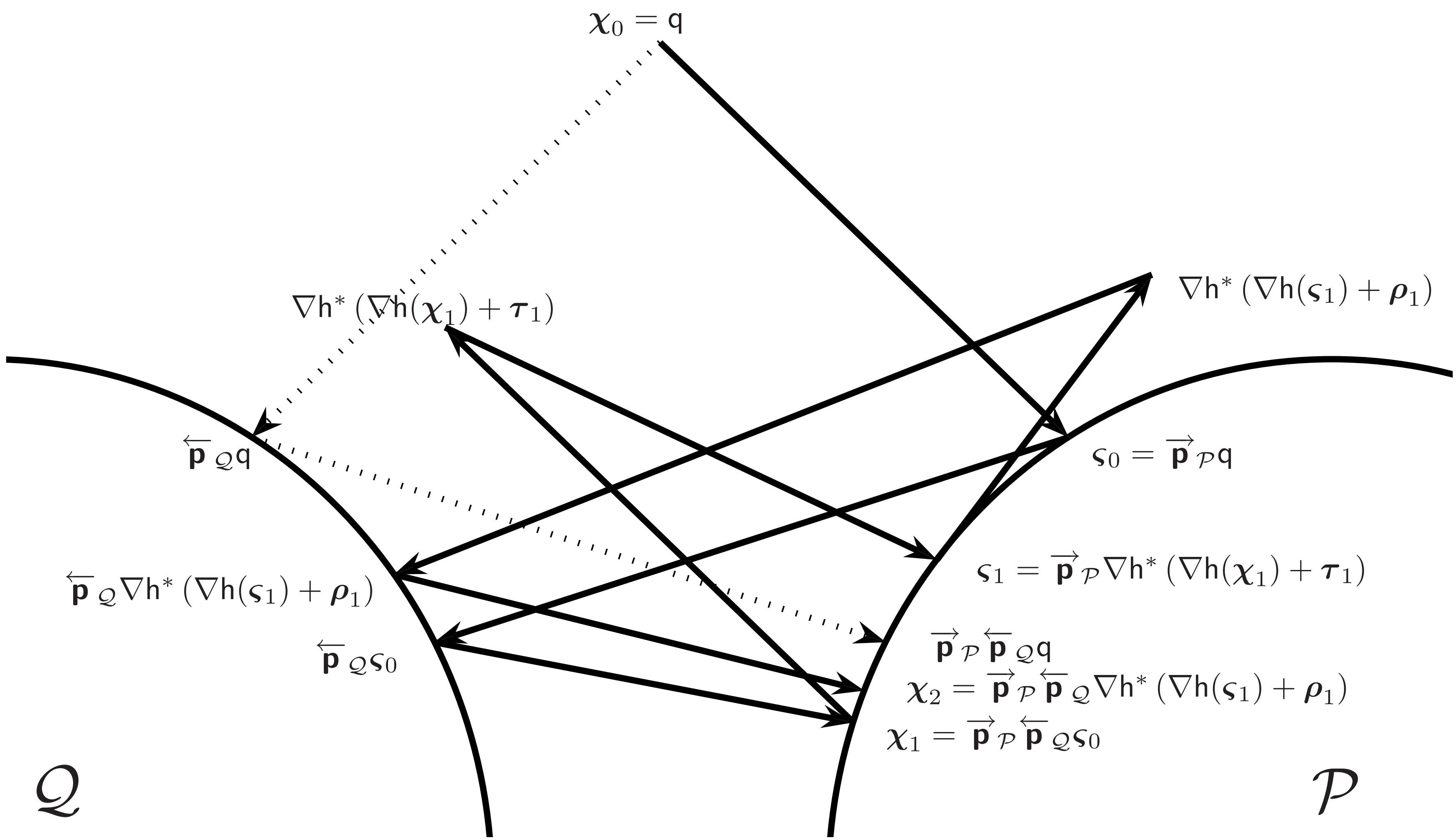
- *Key Result*: Belief Propagation is the Dykstra-like algorithm

$$\begin{aligned}
\boldsymbol{\varsigma}_{\mathrm{k}} &:= \overrightarrow{\mathbf{P}}_{\mathcal{P}} \circ \nabla f^* \left(\nabla f(\boldsymbol{\chi}_{\mathrm{k}}) + \boldsymbol{\tau}_{\mathrm{k}}\right) && (1) \\
\boldsymbol{\tau}_{\mathrm{k}+1} &:= \nabla f(\boldsymbol{\chi}_{\mathrm{k}}) + \boldsymbol{\tau}_{\mathrm{k}} - \nabla f(\boldsymbol{\varsigma}_{\mathrm{k}}) \\
\boldsymbol{\chi}_{\mathrm{k}+1} &:= \overrightarrow{\mathbf{P}}_{\mathcal{P}} \circ \overleftarrow{\mathbf{P}}_{\mathcal{Q}} \circ \nabla f^* \left(\nabla f(\boldsymbol{\varsigma}_{\mathrm{k}}) + \boldsymbol{\rho}_{\mathrm{k}}\right) && (2) \\
\boldsymbol{\rho}_{\mathrm{k}+1} &:= \nabla f(\boldsymbol{\varsigma}_{\mathrm{k}}) + \boldsymbol{\rho}_{\mathrm{k}} - \nabla f(\boldsymbol{\chi}_{\mathrm{k}+1})
\end{aligned}$$

  with f the negative Shannon entropy (i.e. KL projections) and $\boldsymbol{\rho}_0, \boldsymbol{\tau}_0 = \mathbf{0}$.

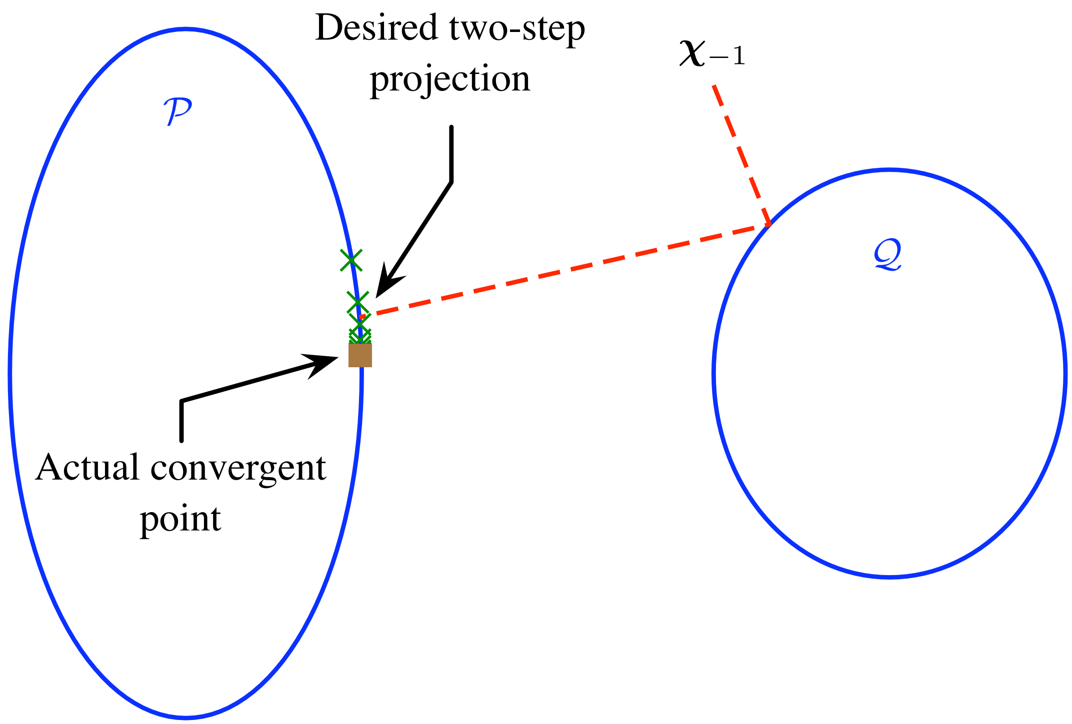
- Hybrid between alternating Bregman projections and Dykstra's algorithm with cyclic Bregman projections.

Belief Propagation as a Hybrid Projections Algorithm



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:

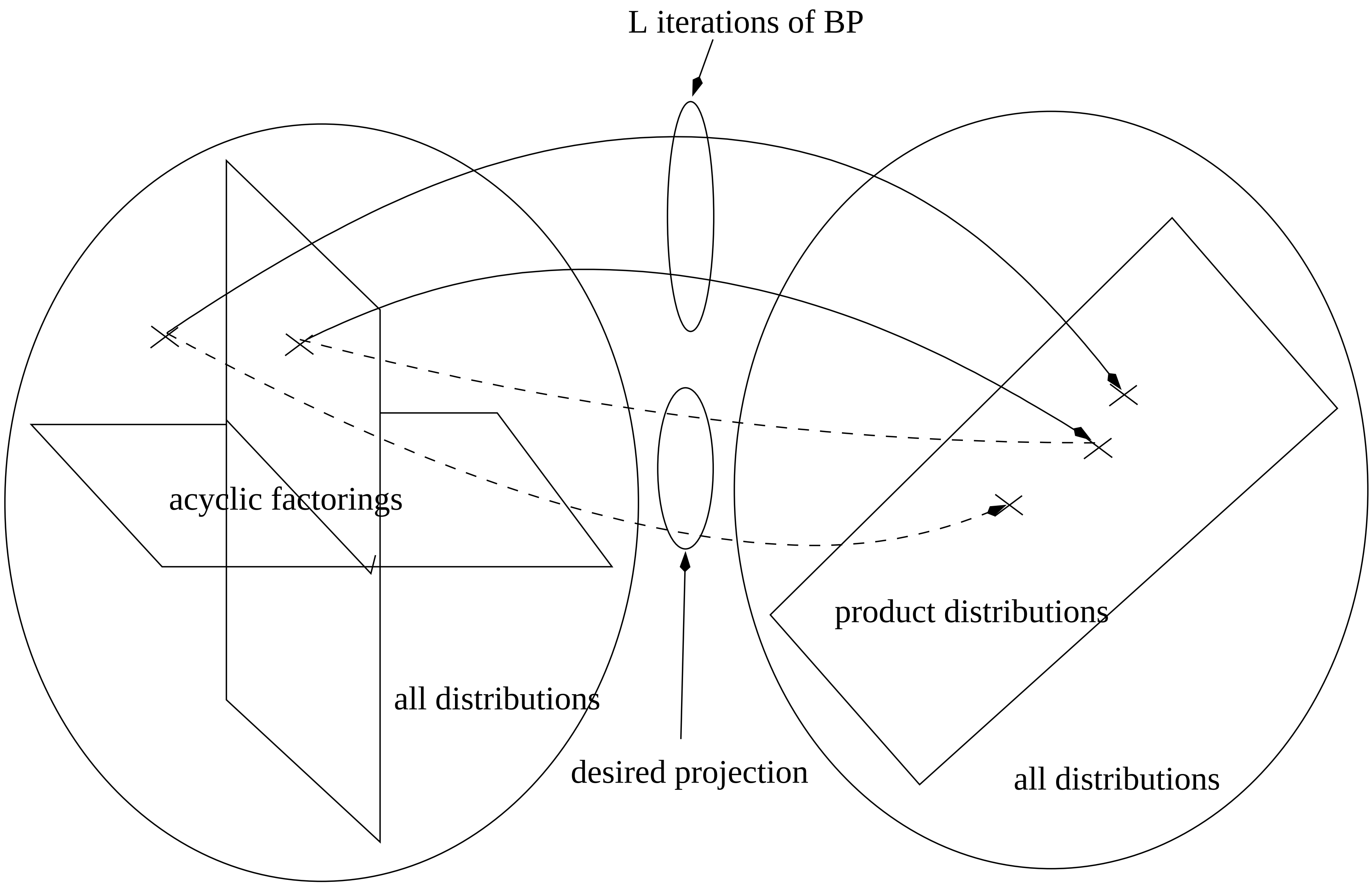
- well known that BP gives correct marginals on acyclic factor graphs (trees and forests)
- how can we translate this to our information geometric framework? (i.e. What region of initializations χ_0 does this correspond to?)
- how can we use the new framework to get a new larger (factor graph girth independent!) 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 χ_0 From Acyclic Factorings

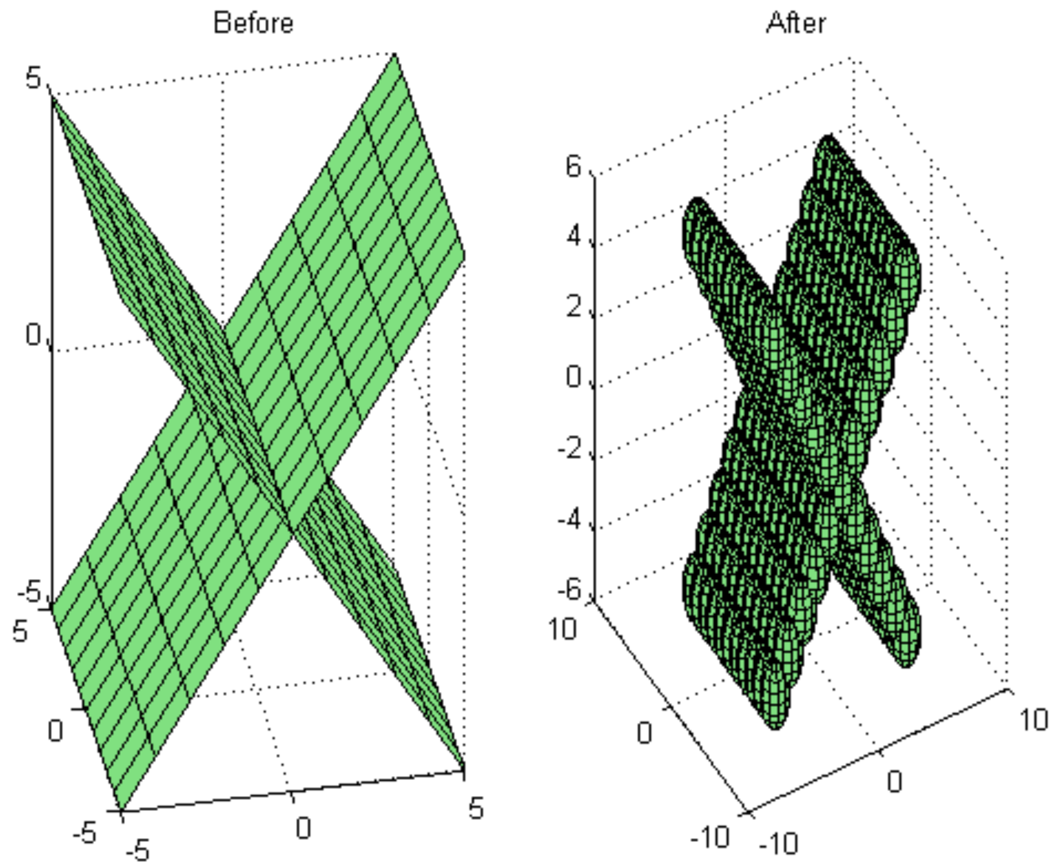
$$\boldsymbol{\chi}_0 = \prod_{k=1}^{K} g_k(\mathbf{x}^k)$$

- Parameterize these by $2^M - 1$ dimensional log coordinates $\boldsymbol{\theta}_k$ of factors $g_k(\mathbf{x}^k)$
- Independence of a factor on a variable $\Leftrightarrow$ $\boldsymbol{\theta}_k$ lies in particular vector subspace.
- So a particular acyclic graph is associated with a particular collection of vector subspaces which $\{\boldsymbol{\theta}_k\}$ 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



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.
  - $\Longrightarrow$ (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

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

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

[3] S. Amari and H. Nagaoka, *Methods of Information Geometry*. AMS Translations of Mathematical Monographs, 2004, vol. 191.

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

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

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

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

[8] L. M. Bregman, "Proof of the convergence of sheleikhovskii's method for a problem with transportation constraints," *USSR Journal on Computational Mathematics and Mathematical Physics*, vol. 7, no. 1, pp. 147–156, 1967.

[9] L. Gubin, B. Polyak, and E. Raik, "The method of projections for finding the common point of convex sets," *USSR Journal on Computational Mathematics and Mathematical Physics*, vol. 7, no. 6, pp. 1211–1228, 1967.

[10] H. Bauschke and J. Borwein, "Legendre functions and the method of random Bregman projections," *Journal of Convex Analysis*, no. 4(1), pp. 27–67, 1997.

[11] H. Bauschke and D. Noll, "The method of forward projections," *Journal of Nonlinear and Convex Analysis*, vol. 3, no. 2, pp. 191–205, 2002.

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

[13] A. J. Grant, "Information geometry and iterative decoding," in *Proceedings IEEE Communication Theory Workshop*, may 1999.

[14] T. J. Richardson, "The geometry of turbo-decoding dynamics." *IEEE Trans. Inform. Theory*, vol. 46, pp. 9–23, Jan. 2000.

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

[16] ——, "Information geometry of turbo and low-density parity-check codes," *IEEE Trans. Inform. Theory*, vol. 50, pp. 1097 – 1114, June 2004.