Information Geometric view of Belief Propagation

Yunshu Liu

2013-10-17

References:

[1]. Shiro Ikeda, Toshiyuki Tanaka and Shun-ichi Amari, *Stochastic reasoning, Free energy and Information Geometry*, Neural Computation, 16, 1779-1810, 2004.

[2]. Shiro Ikeda, Toshiyuki Tanaka and Shun-ichi Amari, *Information Geometry of Turbo and Low-Density Parity-Check Codes*, IEEE Transaction on Information Theory, VOL. 50, NO.6, June 2004.

[3].Kevin P. Murphy, *Machine Learning - A Probabilistic Perspective*, The MIT Press, 2012.

Outline

Probability distribution viewed as manifold

Outline for Infromation Geometry

Examples of probabilistic manifold

Important Submanifold

Information projection

Manifold

Manifold S

Manifold: a set with a coordinate system, a one-to-one mapping from S to $\mathbb{R}^n$, supposed to be "locally" looks like an open subset of $\mathbb{R}^n$"

Elements of the set(points): points in $\mathbb{R}^n$, probability distribution, linear system.

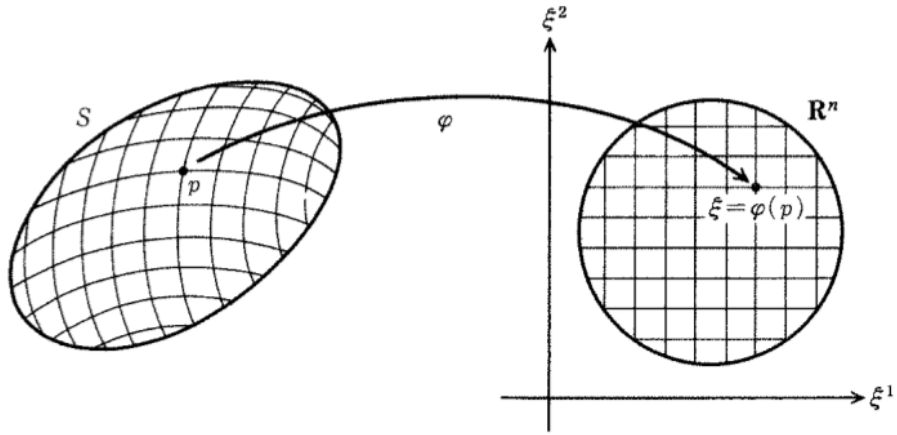


Figure: A coordinate system ξ for a manifold S

Example manifold

Parametrization of color models

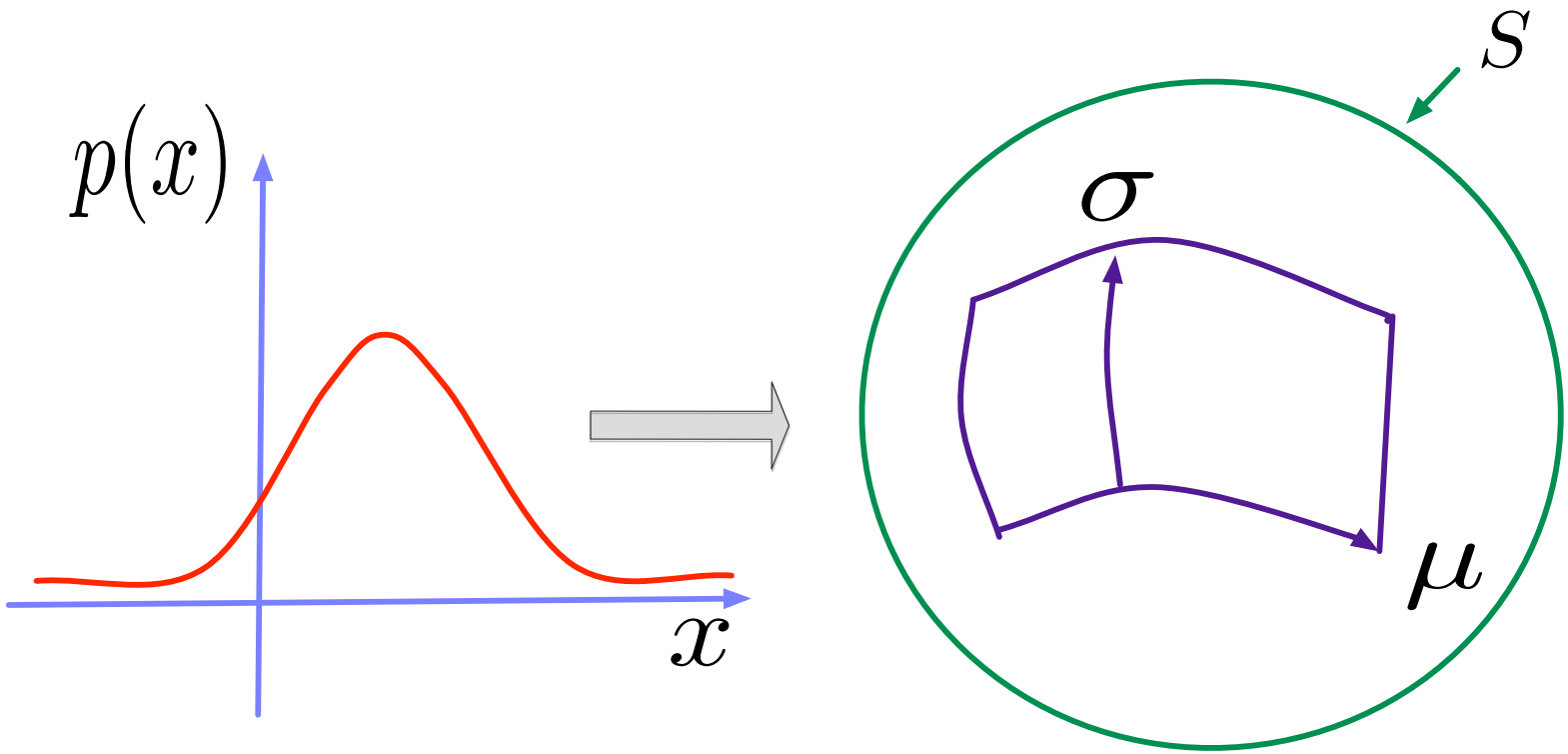
Parametrization: map of color into $\mathbb{R}^3$

Examples:

Example of manifold

Parametrization of Gaussian distribution

$$S = \{p(x;\mu,\sigma)\} \quad p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$$



Example of manifold

Exponential family

$$p(x;\theta) = exp[C(x) + \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta)]$$

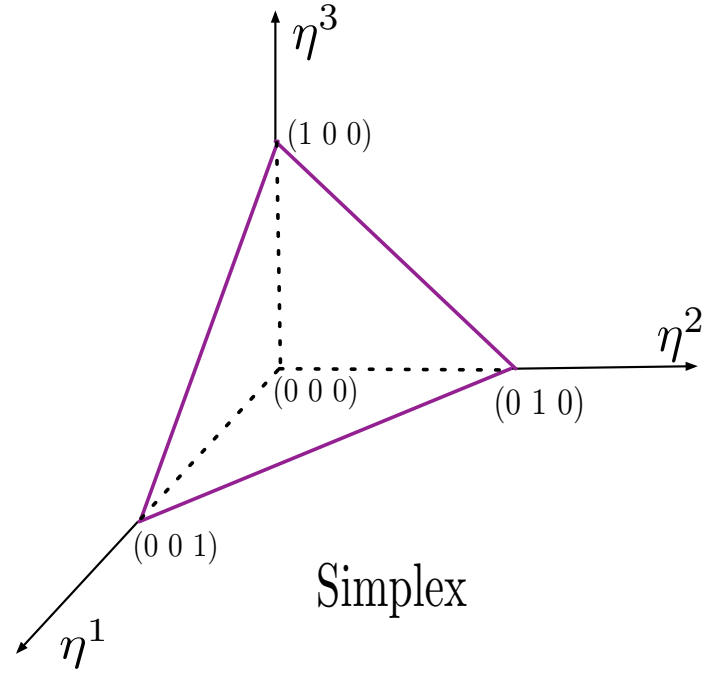
$[\theta^i]$ are called the natural parameters(coordinates), and ψ is the potential function for $[\theta^i]$, which can be calculated as

$$\psi(\theta) = log \int exp[C(x) + \sum_{i=1}^{n} \theta^i F_i(x)]dx$$

The exponential families include many of the most common distributions, including the normal, exponential, gamma, beta, Dirichlet, Bernoulli, binomial, multinomial, Poisson, and so on.

Example of two binary discrete random variables as Exponential families

Let $x = \{x_1, x_2\}$, $x_i = 1$ *or* -1,
consider the family of all the
probability distributions $S =$
$\{p(x) \mid p(x) > 0, \sum_i p(x = C_i) = 1\}$
Outcomes:
$C_0 = \{-1, -1\}, C_1 = \{-1, 1\}$
$C_2 = \{1, -1\}, \ C_3 = \{1, 1\}$
Parameterization:
$p(x = C_i) = \eta^i$ *for* $i = 1, 2, 3$
$p(x = C_0) = 1 - \sum_{j=1}^{3} \eta^j$
$\eta = (\eta^1 \ \eta^2 \ \eta^3)$ is called the
m-coordinate of S.

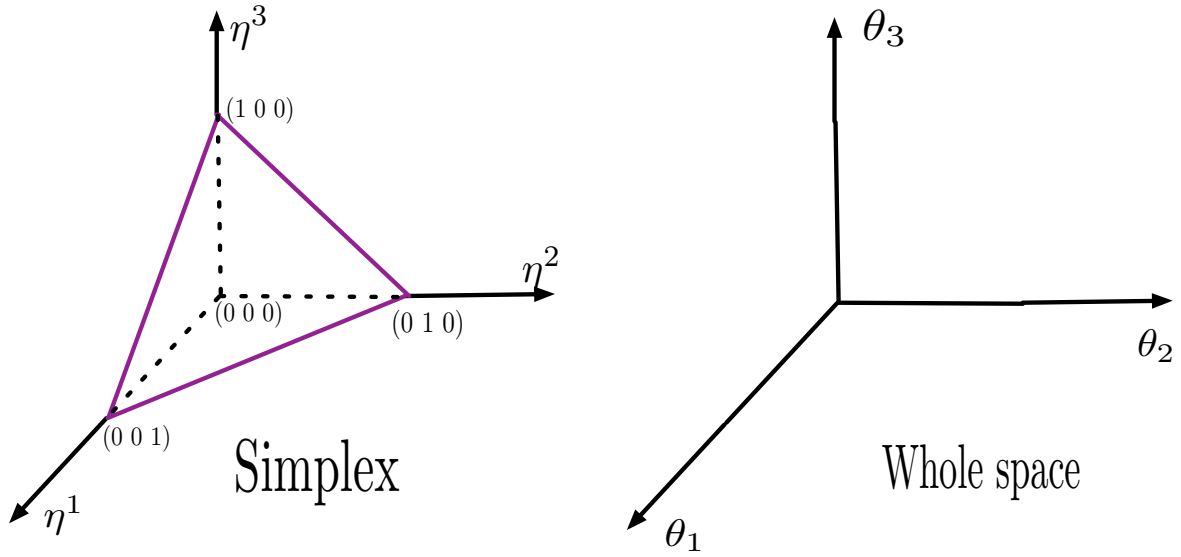


Example of two binary discrete random variables as Exponential families

A new coordinate $\{\theta_i\}$, which is defined as

$$\theta_i = ln \frac{\eta^i}{1 - \sum_{j=1}^{3} \eta^j} \text{ for } i = 1, 2, 3. \quad (1)$$

$\theta = (\theta_1\ \theta_2\ \theta_3)$ is called the e-coordinate of S.



Example of two binary discrete random variables as Exponential families

Introduce new random variables:

$$\delta_{C_i}(x) = \begin{cases} 1 & \text{if x = } C_i \\ 0 & \text{otherwise} \end{cases}$$

and let

$$\psi = -ln\ p(x = C_0) \tag{2}$$

Then S is an exponential family with natural parameter $\{\theta_i\}$:

$$p(x, \theta) = exp(\sum_{i=1}^{3} \theta_i \delta_{C_i}(x) - \psi(\theta)) \tag{3}$$

Probability distribution viewed as manifold

Outline for Infromation Geometry

Examples of probabilistic manifold

Important Submanifolds

Information projection

Submanifolds

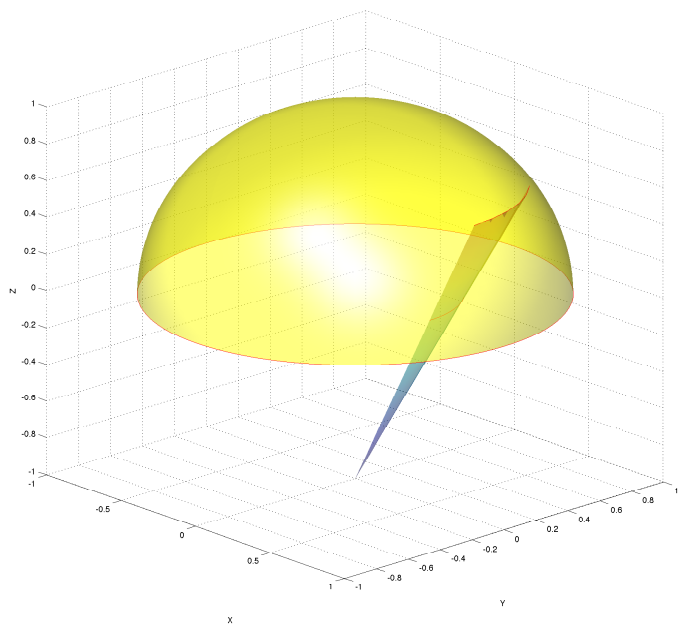
Submanifolds

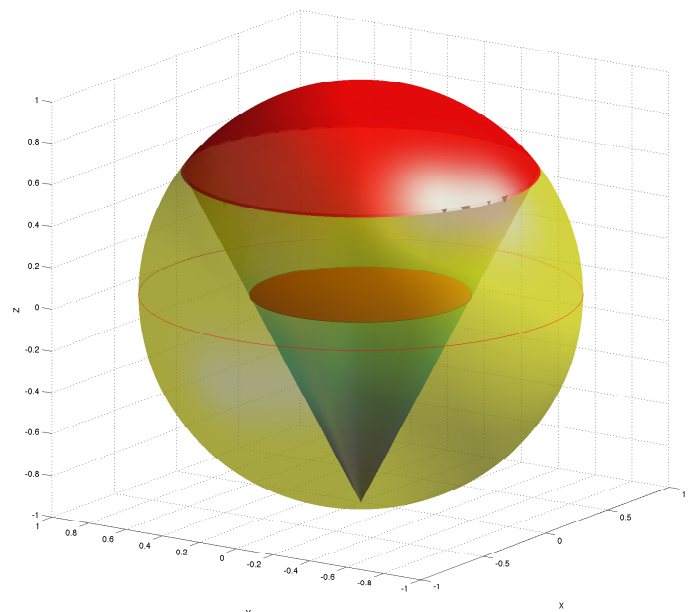
Definition: a submanifold M of a manifold S is a subset of S which itself has the structure of a manifold

An open subset of n-dimensional manifold forms an n-dimensional submanifold.

One way to construct m($<$n)dimensional manifold: fix n-m coordinates.

Examples:





e-autoparallel submanifold and m-autoparallel submanifold

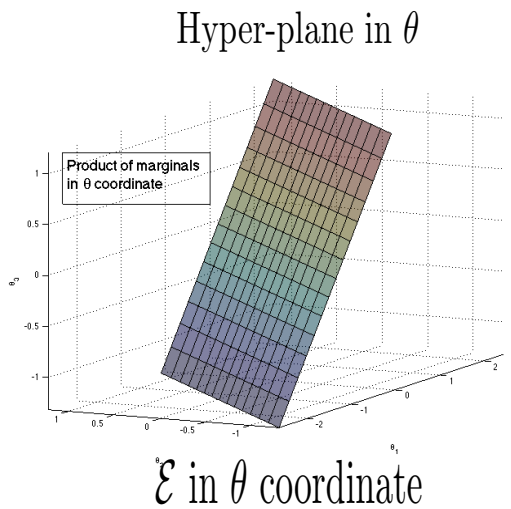
e-autoparallel submanifold

A subfamily E is said to be e-autoparallel submanifold of S if for some coordinate λ there exist a matrix A and a vector b such that for the e-coordinate θ of S

$$\theta(p) = \mathbf{A}\boldsymbol{\lambda}(p) + \mathbf{b} \quad (\forall p \in E) \tag{4}$$

If λ is one-dimensional, it is called a *e-geodesic*.

An e-autoparallel submanifold of S is a hyperplane in e-coordinate;
A e-geodesic of S is a straight line in e-coordinate.



e-autoparallel submanifold and m-autoparallel submanifold

m-autoparallel submanifold

A subfamily M is said to be m-autoparallel submanifold of S if for some coordinate γ there exist a matrix A and a vector b such that for the m-coordinate η of S

$$\boldsymbol{\eta}(p) = \mathbf{A}\boldsymbol{\gamma}(p) + \mathbf{b} \quad (\forall p \in M) \tag{5}$$

If γ is one-dimensional, it is called a *m-geodesic*.

e-autoparallel submanifold and m-autoparallel submanifold

Two independent bits

The set of all product distributions, which is defined as

$$\mathcal{E} = \{p(x) \mid p(x) = p_{X_1}(x_1)p_{X_2}(x_2)\} \tag{6}$$

$p_{X_1}(x_1 = 1) = p_1$, $p_{X_2}(x_2 = 1) = p_2$
then $p_{X_1}(x_1 = -1) = 1 - p_1$, $p_{X_2}(x_2 = -1) = 1 - p_2$

Parameterization

Define $\lambda_i = \frac{1}{2}\, ln\, \frac{p_i}{1-p_i}$ for i = 1, 2, and
$\psi(\lambda) = \sum_{i=1,2} ln\, (e^{\lambda_i} + e^{-\lambda_i})$,

$$p(x, \lambda) = e^{\lambda_1 x_1 - \psi(\lambda_1)} e^{\lambda_2 x_2 - \psi(\lambda_2)} = e^{\lambda x - \psi(\lambda)} \tag{7}$$

Thus $\mathcal{E}$ is an exponential family with nature parameter $\{\lambda_i\}$

e-autoparallel submanifold and m-autoparallel submanifold

Dual parameters: expectations

$$\gamma = \begin{pmatrix} \gamma^1 \\ \gamma^2 \end{pmatrix} = \begin{pmatrix} \partial_1 \psi(\lambda) \\ \partial_2 \psi(\lambda) \end{pmatrix} = \begin{pmatrix} 2p_1 - 1 \\ 2p_2 - 1 \end{pmatrix} = E_p(x)$$

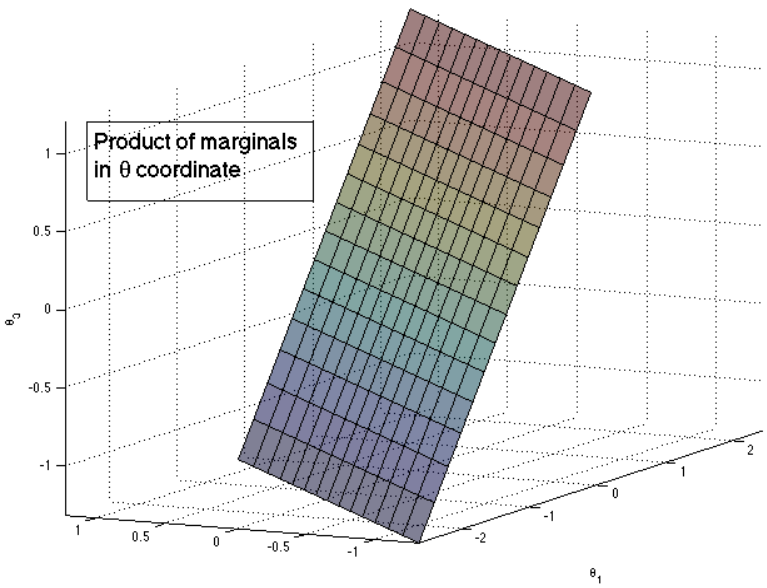
Properties

$\mathcal{E}$ is an e-autoparallel submanifold(plane in θ coordinate) of S, but not a m-autoparallel submanifold(not a plane in η coordinate):

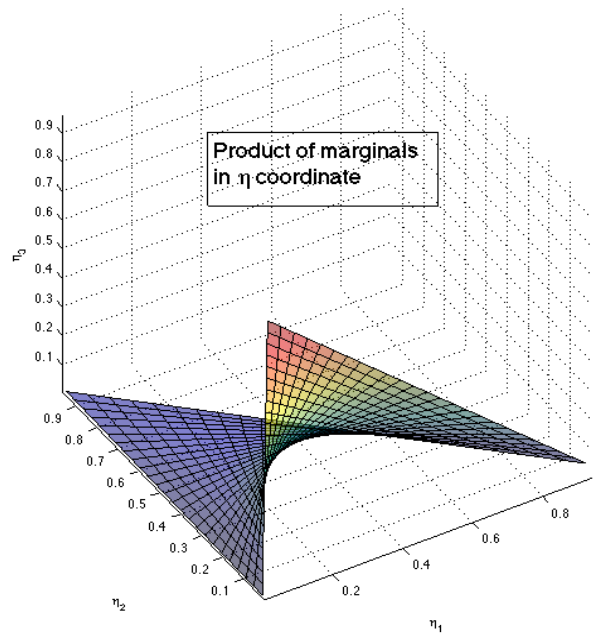
$$\theta(p) = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 0 \\ 0 & 2 \end{pmatrix} \lambda(p)$$

e-autoparallel submanifold and m-autoparallel submanifold

Two independent bits: e-autoparallel submanifold but not m-autoparallel submanifold



$\mathcal{E}$ in θ coordinate



$\mathcal{E}$ in η coordinate

Probability distribution viewed as manifold

Outline for Infromation Geometry

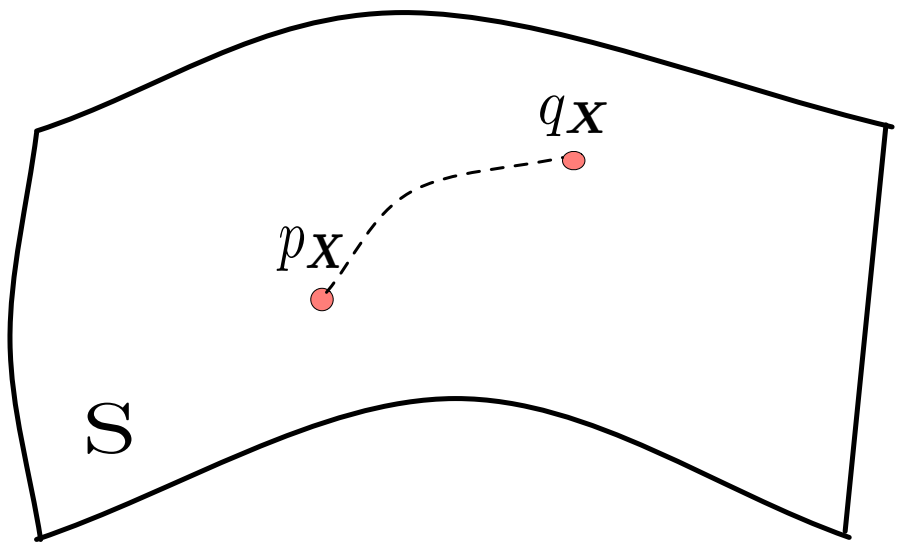
Examples of probabilistic manifold
Important Submanifolds
Information projection

Projections and Pythagorean Theorem

KL-divergence

On the manifold of probability mass functions for random variables taking values in the set $\mathcal{X}$, we can also define the Kullback Leibler divergence, or relative entropy, measured in bits, according to

$$D(p_{\boldsymbol{X}}||q_{\boldsymbol{X}}) = \sum_{\boldsymbol{x}\in\mathcal{X}} p_{\boldsymbol{X}}(\boldsymbol{x}) \log\left(\frac{p_{\boldsymbol{X}}(\boldsymbol{x})}{q_{\boldsymbol{X}}(\boldsymbol{x})}\right) \tag{8}$$



Projections and Pythagorean Theorem

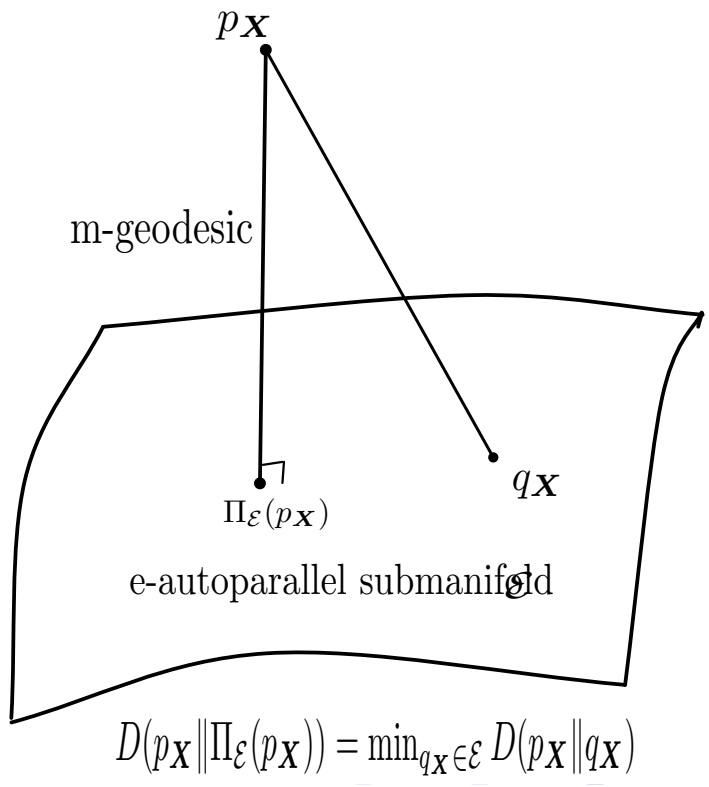
Information Projection

Let p be a point in S and let $\mathcal{E}$ be a e-autoparallel submanifold of S. We want to find point $\Pi_{\mathcal{E}}(p_{\boldsymbol{X}})$ in $\mathcal{E}$ that minimize the KL-divergence

$$D(p_{\boldsymbol{X}} \| \Pi_{\mathcal{E}}(p_{\boldsymbol{X}})) = \min_{q_{\boldsymbol{X}} \in \mathcal{E}} D(p_{\boldsymbol{X}} \| q_{\boldsymbol{X}}) \tag{9}$$

The m-geodesic connecting $p_{\boldsymbol{X}}$ and $\Pi_{\mathcal{E}}(p_{\boldsymbol{X}})$ is orthogonal to $\mathcal{E}$ at $\Pi_{\mathcal{E}}(p_{\boldsymbol{X}})$.

- In geometry, orthogonal usually means right angle
- In information geometry, orthogonal usually means certain values are uncorrelated

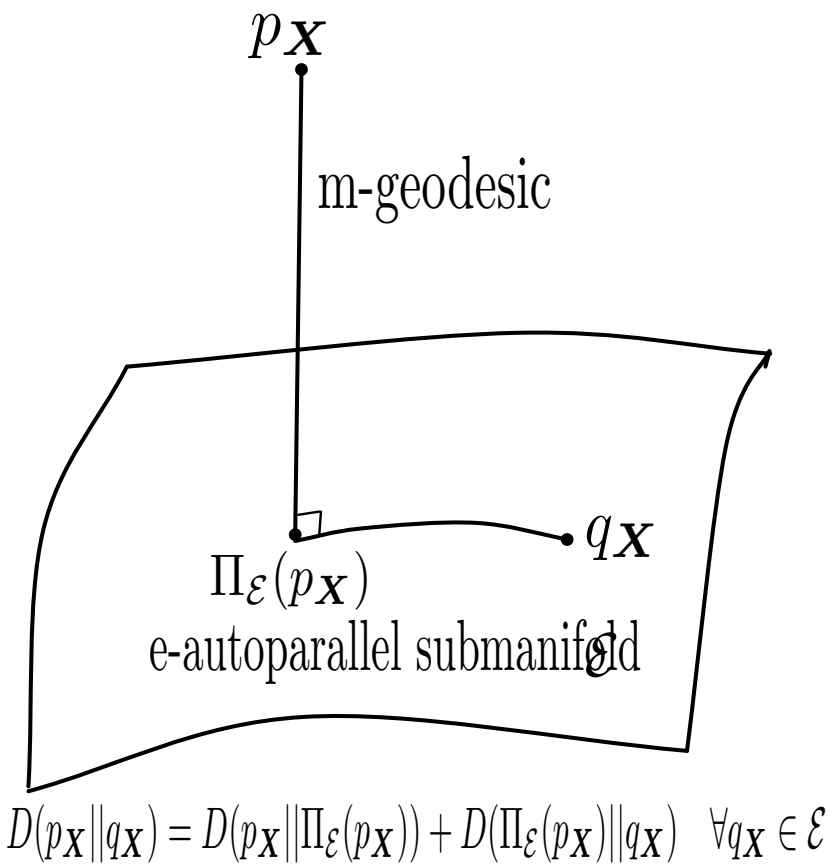


Projections and Pythagorean Theorem

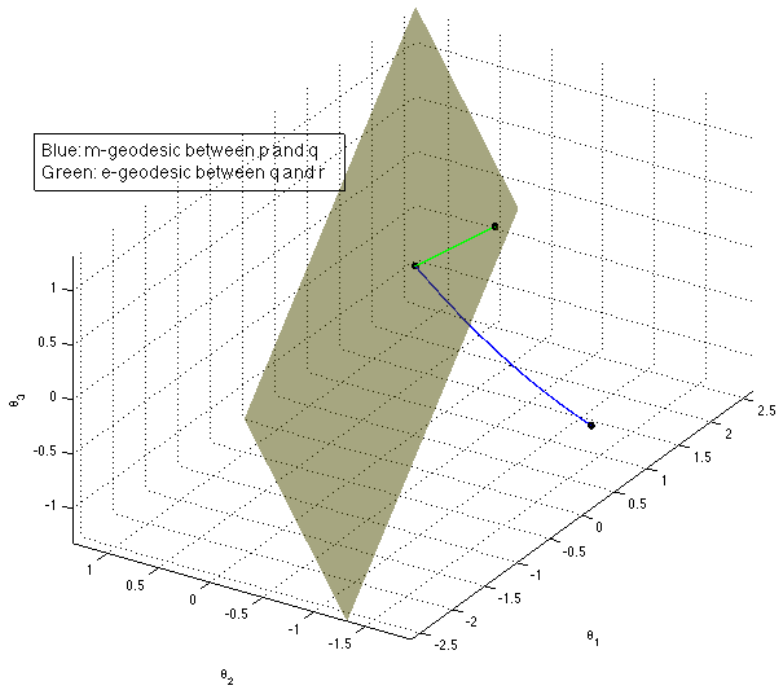
Pythagorean relation

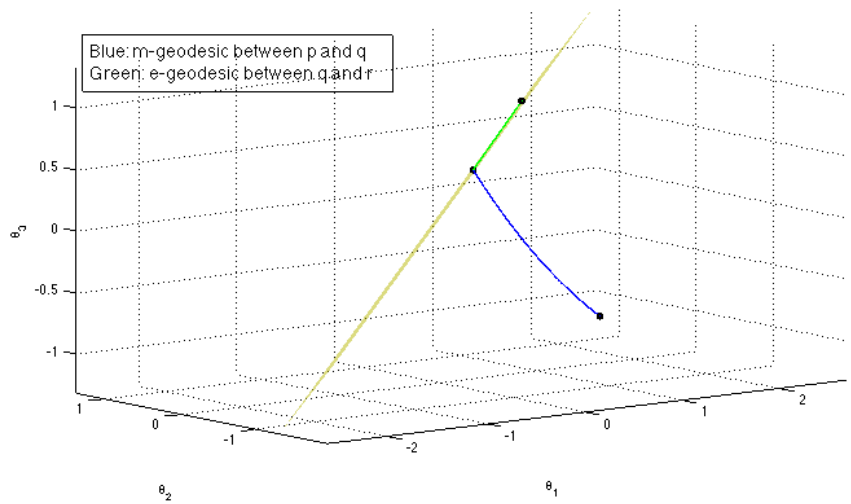
For $\forall q_{\boldsymbol{X}} \in \mathcal{E}$, we have the following Pythagorean relation.

$$D(p_{\boldsymbol{X}}||q_{\boldsymbol{X}}) = D(p_{\boldsymbol{X}}||\Pi_{\mathcal{E}}(p_{\boldsymbol{X}})) + D(\Pi_{\mathcal{E}}(p_{\boldsymbol{X}})||q_{\boldsymbol{X}}) \quad \forall q_{\boldsymbol{X}} \in \mathcal{E} \quad (10)$$

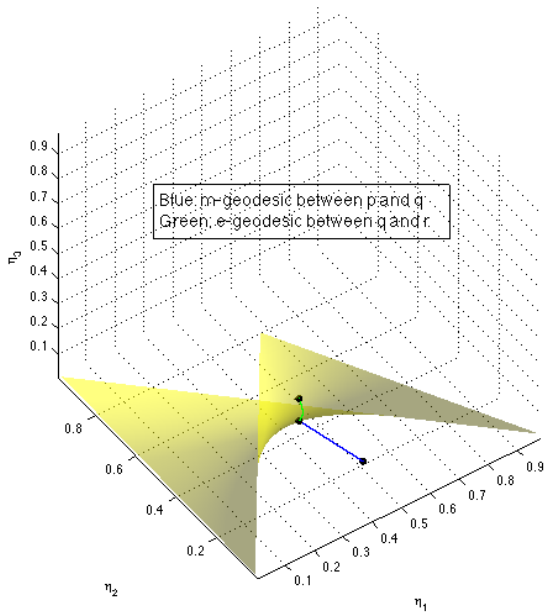


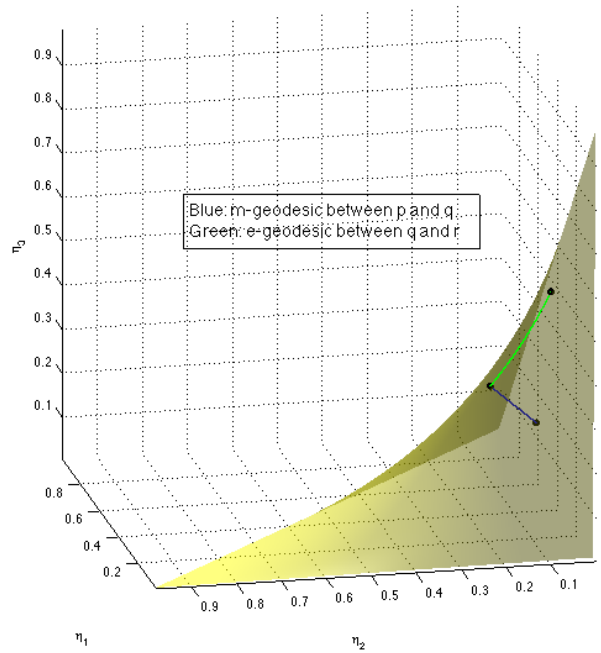
Projection in θ coordinate





Projection in η coordinate





Information Geometry of Belief Propagation

Outline for Information Geometry of Belief Propagation

Belief Propagation/Sum-product algorithms

Information Geometry of Graphical structure

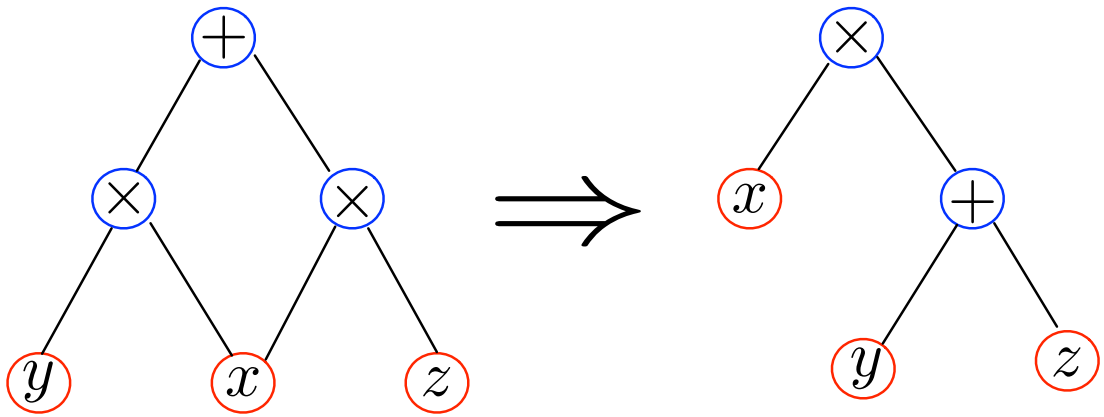
Information-Geometrical view of Belief Propagation

Belief Propagation

Belief Propagation/Sum-product algorithm

The Belief Propagation algorithm involves using a message passing scheme to explore the conditional independence and use these properties to change the order of sum and product.

$$xy + xz = x(y + z)$$



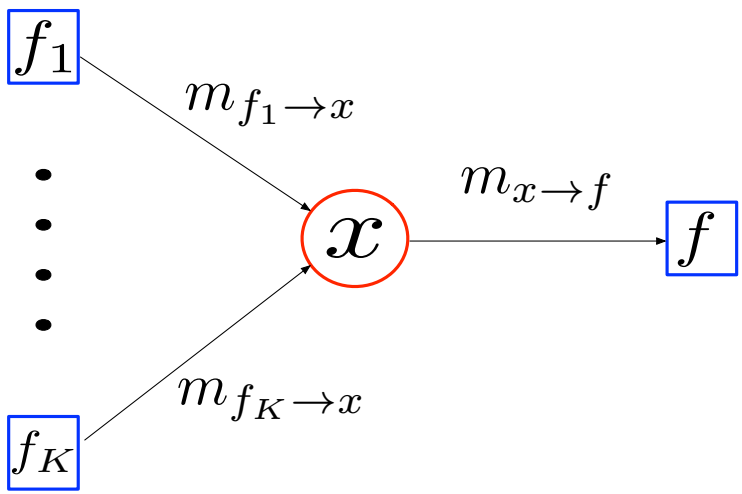
Expression tree of xy + xz and x(y+z), where the leaf nodes represent variables and internal nodes represent arithmetic operators(addition, multiplication, negation, etc.).

Belief Propagation

Message passing on a factor graph:

Variable x to factor f:

$$m_{x \to f} = \prod_{k \in nb(x) \setminus f} m_{f_k \to x}(x)$$



$nb(x)$ denote the neighboring factor nodes of x

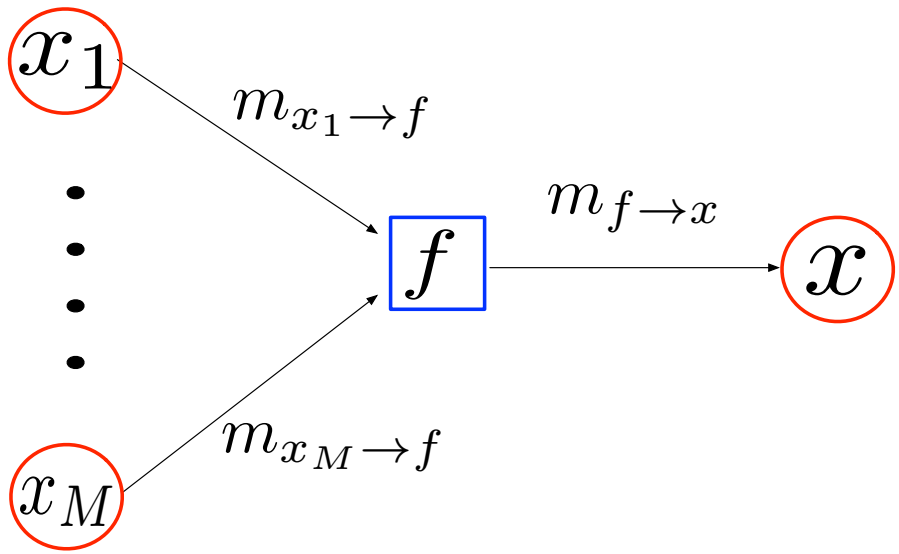
If $nb(x) \setminus f = \emptyset$ then $m_{x \to f} = 1$.

Belief Propagation

Message passing on a factor graph:

Factor f to variable x:

$$m_{f\to x} = \sum_{x_1,\cdots x_M} f(x, x_1, \cdots, x_M) \prod_{m\in nb(f)\backslash x} m_{x_m\to f}(x_m)$$



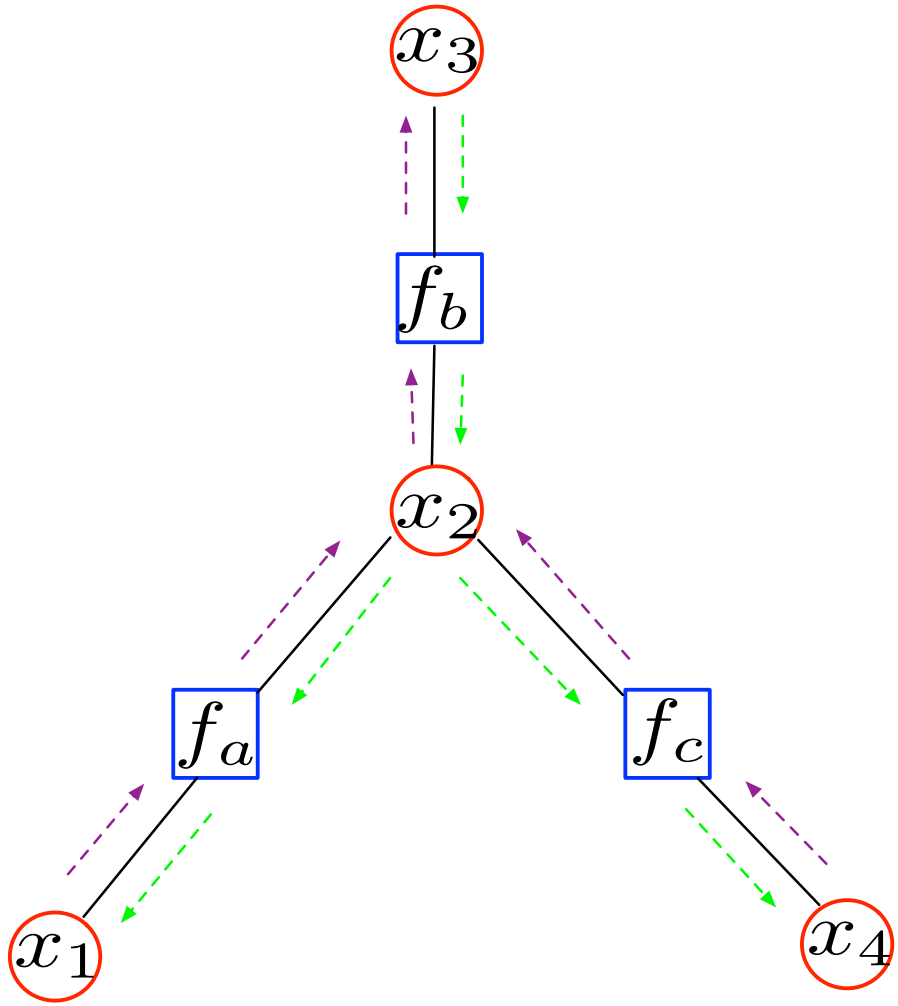
$nb(f)$ denote the neighboring variable nodes of f

If $nb(f)\backslash x = \emptyset$ then $m_{f\to x} = f(x)$.

Serial protocol of Belief Propagation

Message passing on a tree structured factor graph

Serial protocol: Send messages up to the root and back



Serial protocol of Belief Propagation

Message passing on a tree structured factor graph

Down - up:

$$
\begin{aligned}
m_{x_1 \to f_a}(x_1) &= 1 \\
m_{f_a \to x_2}(x_2) &= \sum_{x_1} f_a(x_1, x_2) \\
m_{x_4 \to f_c}(x_4) &= 1 \\
m_{f_c \to x_2}(x_2) &= \sum_{x_4} f_c(x_2, x_4) \\
m_{x_2 \to f_b}(x_2) &= m_{f_a \to x_2}(x_2) m_{f_c \to x_2}(x_2) \\
m_{f_b \to x_3}(x_3) &= \sum_{x_2} f_b(x_2, x_3) m_{x_2 \to f_b}
\end{aligned}
$$

Serial protocol of Belief Propagation

Message passing on a tree structured factor graph

Up - down:

$$
\begin{aligned}
m_{x_3 \to f_b}(x_3) &= 1 \\
m_{f_b \to x_2}(x_2) &= \sum_{x_3} f_b(x_2, x_3) \\
m_{x_2 \to f_a}(x_2) &= m_{f_b \to x_2}(x_2) m_{f_c \to x_2}(x_2) \\
m_{f_a \to x_1}(x_1) &= \sum_{x_2} f_a(x_1, x_2) m_{x_2 \to f_a}(x_2) \\
m_{x_2 \to f_c}(x_2) &= m_{f_a \to x_2}(x_2) m_{f_b \to x_2}(x_2) \\
m_{f_c \to x_4}(x_4) &= \sum_{x_2} f_c(x_2 . x_4) m_{x_2 \to f_c}(x_2)
\end{aligned}
$$

Serial protocol of Belief Propagation

Message passing on a tree structured factor graph

Calculating $p(x_2)$

$$
\begin{aligned}
p(x_2) &= \frac{1}{Z} m_{f_a \to x_2}(x_2) m_{f_b \to x_2}(x_2) m_{f_c \to x_2}(x_2) \\
&= \frac{1}{Z} \Big\{ \sum_{x_1} f_a(x_1, x_2) \Big\} \Big\{ \sum_{x_3} f_b(x_2, x_3) \Big\} \Big\{ \sum_{x_4} f_c(x_2, x_4) \Big\} \\
&= \frac{1}{Z} \sum_{x_1} \sum_{x_3} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4) \\
&= \sum_{x_1} \sum_{x_3} \sum_{x_4} p(\mathbf{x})
\end{aligned}
$$

Parallel protocol of Belief Propagation

Message passing on general graphs with loops

Parallel protocol:

- Initialize all messages
- All nodes receive message from their neighbors in parallel and update their belief states
- Send new messages back out to their neighbors
- Repeats the above two process until convergence

Information Geometry of Belief Propagation

Outline for Information Geometry of Belief Propagation

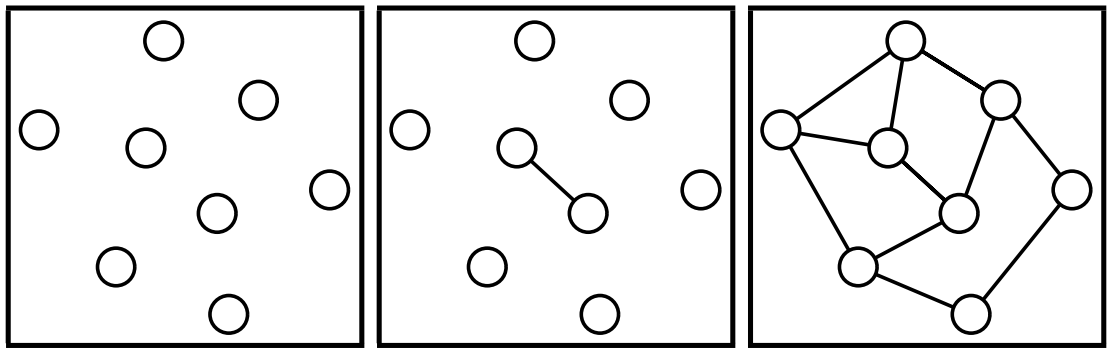
Belief Propagation/Sum-product algorithms

Information Geometry of Graphical structure

Information-Geometrical view of Belief Propagation

Information Geometry of Graphical structure

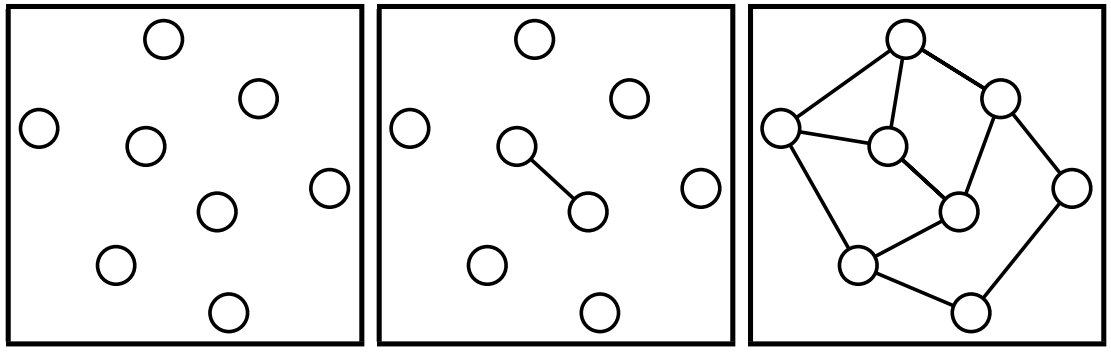
Important Graphs(Undirected) for Belief Propagation



A: Belief graph with no edges(all n nodes are independent)
B: Graph with a single edge r
C: Graph of the true distribution with n nodes and L edges

Information Geometry of Graphical structure

Important Graphs(Undirected) for Belief Propagation



$$A: \quad M_0 = \{p_0(\mathbf{x},\theta) = \exp[\sum_{i=1}^{n} h_i x_i + \sum_{i=1}^{n} \theta_i x_i - \psi_0(\theta)] | \theta \in \mathcal{R}^n\}$$

$$B: \quad M_r = \{p_r(\mathbf{x},\zeta_\mathbf{r}) = \exp[\sum_{i=1}^{n} h_i x_i + s_r c_r(\mathbf{x}) + \sum_{i=1}^{n} \zeta_{ri} x_i - \psi_r(\zeta_\mathbf{r})] | \zeta_\mathbf{r} \in \mathcal{R}^n\}$$

$$C: \quad S = \{p(\mathbf{x},\theta,\mathbf{s}) = exp[\sum_{i=1}^{n} h_i x_i + \sum_{r=1}^{L} s_r c_r(\mathbf{x}) - \psi(\theta,\mathbf{s})] | \theta \in \mathcal{R}^n, \mathbf{s} \in \mathcal{R}^L\}$$

Information Geometry of Graphical structure

Properties of M_0, M_r

- M_0 is an e-autoparallel submanifold of S, with e-affine coordinate θ
- M_r for $r = 1, \cdots L$ are all e-autoparallel submanifold of S, with e-affine coordiante $\zeta_{\mathbf{r}}$
- A distribution in M_r include only one nonlinear interaction term $c_r(\mathbf{x})$ corresponding to the single edge in B

Information Geometry of Belief Propagation

Outline for Information Geometry of Belief Propagation

Belief Propagation/Sum-product algorithms
Information Geometry of Graphical structure
Information-Geometrical view of Belief Propagation

Information-Geometrical view of Belief Propagation

Problem setup for belief propagation

Suppose we have the true distribution $q(\mathbf{x}) \in S$, and want to calculate $q_0(\mathbf{x}) = \prod_{M_0} q(\mathbf{x})$, the m-projection of $q(\mathbf{x})$ to M_0, the goal of belief propagation is to find a good approximation $\hat{q}(\mathbf{x}) \in M_0$ of $q(\mathbf{x})$.

$$M_0 = \{p_0(\mathbf{x}, \theta) = \exp[\sum_{i=1}^{n} h_i x_i + \sum_{i=1}^{n} \theta_i x_i - \psi_0(\theta)] | \theta \in \mathcal{R}^n\}$$

$$M_r = \{p_r(\mathbf{x}, \zeta_\mathbf{r}) = \exp[\sum_{i=1}^{n} h_i x_i + s_r c_r(\mathbf{x}) + \sum_{i=1}^{n} \zeta_{ri} x_i - \psi_r(\zeta_\mathbf{r})] | \zeta_\mathbf{r} \in \mathcal{R}^n\}$$

$$S = \{p(\mathbf{x}, \theta, \mathbf{s}) = exp[\sum_{i=1}^{n} h_i x_i + \sum_{r=1}^{L} s_r c_r(\mathbf{x}) - \psi(\theta, \mathbf{s})] | \theta \in \mathcal{R}^n, \mathbf{s} \in \mathcal{R}^L\}$$

Geometrical Belief Propagation Algorithm

- 1) Put $t = 0$, and start with initial guesses $\zeta_{\mathbf{r}}^{\mathbf{0}}$, for example, $\zeta_{\mathbf{r}}^{\mathbf{0}} = 0$
- 2) For $t = 0, 1, 2, \cdots$, m-project $p_r(\mathbf{x}, \zeta_{\mathbf{r}}^{\mathbf{t}})$ to M_0 and obtain the linearized version of $s_r c_r(\mathbf{x})$

$$\xi_{\mathbf{r}}^{\mathbf{t+1}} = \prod_{M_0} p_r(\mathbf{x}, \zeta_{\mathbf{r}}^{\mathbf{t}}) - \zeta_{\mathbf{r}}^{\mathbf{t}}$$

- 3) Summarize all the effects of $s_r c_r(\mathbf{x})$, to give

$$\theta^{\mathbf{t+1}} = \sum_{r=1}^{L} \xi_{\mathbf{r}}^{\mathbf{t+1}}$$

- 4) Update $\zeta_{\mathbf{r}}$ by

$$\zeta_{\mathbf{r}}^{\mathbf{t+1}} = \sum_{r' \neq r} \xi_{\mathbf{r}'}^{\mathbf{t+1}} = \theta^{\mathbf{t+1}} - \xi_{\mathbf{r}}^{\mathbf{t+1}}$$

- 5) Repeat 2-4 until convergence

Geometrical Belief Propagation Algorithm

Analysis

Given $q(\mathbf{x})$, let $\zeta_{\mathbf{r}}$ be the current solution which M_r believes to give a good approximation of $q(\mathbf{x})$. then we project it to M_0, giving the solution $\theta_{\mathbf{r}}$ in M_0 having the same expectation: $p(\mathbf{x}, \theta_{\mathbf{r}}) = \prod_{M_0} p_r(\mathbf{x}, \zeta_{\mathbf{r}})$.

Since $\theta_{\mathbf{r}}$ includes both the effect of the single nonlinear term $s_r c_r(\mathbf{x})$ specific to M_r and that due to the linear term $\zeta_{\mathbf{r}}$, we can use $\xi_{\mathbf{r}} = \theta_{\mathbf{r}} - \zeta_{\mathbf{r}}$ to represents the effect of the single nonlinear term $s_r c_r(\mathbf{x})$, we consider it as the linearized version of $s_r c_r(\mathbf{x})$.

Geometrical Belief Propagation Algorithm

Analysis:continue

Once M_r knows the linearized version of $s_r c_r(\mathbf{x})$ is $\xi_{\mathbf{r}}$, it broadcasts $\xi_{\mathbf{r}}$ to all the other models M_0 and $M_{r'}(r' \neq r)$. Receiving these messages from all M_r, M_0 guesses that the equivalent linear term of $\sum_{r=1}^{L} s_r c_r(\mathbf{x})$ is $\theta = \sum_{r=1}^{L} \xi_{\mathbf{r}}$.

Since M_r in turn receives messages $\xi_{\mathbf{r}'}$ from all other $M_{r'}(r' \neq r)$, M_r uses them to form a new $\zeta_{\mathbf{r}} = \sum_{r' \neq r} \xi_{\mathbf{r}'} = \theta - \xi_{\mathbf{r}}$. The whole process is repeated until converge.

Geometrical Belief Propagation Algorithm

Analysis:continue

Assume the algorithm has converged to $\{\zeta_{\mathbf{r}}^*\}$ and $\{\theta^*\}$. We find the estimation of $q(\mathbf{x}) \in S$ in M_0 is $p_0(\mathbf{x}, \theta^*)$(which may not be the exact solution). We write $q(\mathbf{x})$, $p_0(\mathbf{x}, \theta^*)$ and $p_r(\mathbf{x}, \zeta_{\mathbf{r}}^*)$ as:

$$
\begin{aligned}
p_0(\mathbf{x}, \theta^*) &= \exp[\sum_{i=1}^{n} h_i x_i + \sum_{r=1}^{L} \xi_{\mathbf{r}}^* \cdot \mathbf{x} - \psi_0] \\
p_r(\mathbf{x}, \zeta_{\mathbf{r}}^*) &= \exp[\sum_{i=1}^{n} h_i x_i + s_r c_r(\mathbf{x}) + \sum_{r' \neq r} \xi_{\mathbf{r}'}^* \cdot \mathbf{x} - \psi_r] \\
q(\mathbf{x}) &= exp[\sum_{i=1}^{n} h_i x_i + \sum_{r=1}^{L} s_r c_r(\mathbf{x}) - \psi]
\end{aligned}
$$

The whole idea of the algorithm is to approximate $s_r c_r(\mathbf{x})$ by $\xi_{\mathbf{r}}^* \cdot \mathbf{x}$ in M_r, then the independent distribution $p_0(\mathbf{x}, \theta^*) \in M_0$ integrates all the information.

Information-Geometrical view of Belief Propagation

Convergence Analysis

The convergence point satisfies the two conditions:

- 1. m-condition: $\theta^* = \prod_{M_0} p_r(\mathbf{x}, \zeta^*_{\mathbf{r}})$
- 2. e-condition: $\theta^* = \frac{1}{L-1} \sum_{r=1}^{L} \zeta^*_{\mathbf{r}}$

The m-condition and e-condition of the algorithm

- e-condition is satisfied at each iteration of BP algorithm
- m-condition is satisfied at the equilibrium of the BP algorithm

Information-Geometrical view of Belief Propagation

Convergence Analysis

Let us consider the m-autoparallel submanifold M^* connecting $p_0(\mathbf{x}, \theta^*)$ and $p_r(\mathbf{x}, \zeta_{\mathbf{r}}^*), r = 1, \cdots, L$:

$$M^* = \{p(\mathbf{x}) | p(\mathbf{x}) = d_0 p_0(\mathbf{x}, \theta^*) + \sum_{r=1}^{L} d_r p_r(\mathbf{x}, \zeta_{\mathbf{r}}^*); d_0 + \sum_{r=1}^{L} d_r = 1\}$$

We also consider the e-autoparallel submanifold E^* connecting $p_0(\mathbf{x}, \theta^*)$ and $p_r(\mathbf{x}, \zeta_{\mathbf{r}}^*)$:

$$E^* = \{p(\mathbf{x}) | \log p(\mathbf{x}) = v_0 \log p_0(\mathbf{x}, \theta^*) + \sum_{r=1}^{L} v_r \log p_r(\mathbf{x}, \zeta_{\mathbf{r}}^*) - \psi;$$

$$v_0 + \sum_{r=1}^{L} v_r = 1\}$$

Information-Geometrical view of Belief Propagation

Convergence Analysis

M^* hold the expectation of $\mathbf{x}$ unchange, thus a equivalent definition is given by:

$$M^* = \{p(\mathbf{x}) | p(\mathbf{x}) \in S, \sum_{\mathbf{x}} \mathbf{x} p(\mathbf{x}) = \sum_{\mathbf{x}} \mathbf{x} p_0(\mathbf{x}, \theta^*) = \eta_{\mathbf{0}}(\theta^*)\}$$

Thus m-condition implies that M^* intersects all of M_0 and M_r orthogonally.

The e-condition implies that E^* include the true distribution $q(\mathbf{x})$ since if we set $v_0 = -(L-1)$, $v_r = 1$ for $\forall r$, using e-condition, we will get

$$q(\mathbf{x}) = C p_0(\mathbf{x}, \theta^*)^{-(L-1)} \prod_{r=1}^{L} p_r(\mathbf{x}, \zeta_{\mathbf{r}}^*) \in E^*$$

where C is the normalization factor

Information-Geometrical view of Belief Propagation

Convergence Analysis

The proposed BP algorithm search for θ^* and ζ^* until both the m-condition and e-condition are satisfied. We know E^* includes $q(\mathbf{x})$, if M^* include $q(\mathbf{x})$, its m-projection to M_0 is $p_0(\mathbf{x}, \theta^*)$, θ^* gives the true solution; if M^* doesn't include $q(\mathbf{x})$, the algorithm only give a approximation of the true distribution. Thus we have the following theorem:

Theorem: When the underlying graph is a tree, both E^* and M^* include $q(\mathbf{x})$, and the algorithm gives the exact solution.

Information-Geometrical view of Belief Propagation

Example of L = 2 where $\theta = \zeta_1 + \zeta_2 = \xi_1 + \xi_2$, $\zeta_1 = \xi_2$ and $\zeta_2 = \xi_1$

- 1) Put $t = 0$, and start with initial guesses ζ_2^0, for example, $\zeta_2^0 = 0$
- 2) For $t = 0, 1, 2, \cdots$, m-project $p_r(\mathbf{x}, \zeta_r^t)$ to M_0 and obtain the linearized version of $s_r c_r(\mathbf{x})$

$$\zeta_1^{t+1} = \prod_{M_0} p_2(\mathbf{x}, \zeta_2^t) - \zeta_2^t, \quad \zeta_2^{t+1} = \prod_{M_0} p_1(\mathbf{x}, \zeta_1^{t+1}) - \zeta_1^{t+1}$$

- 3)Summarize all the effects of $s_r c_r(\mathbf{x})$, to give

$$\theta^{t+1} = \zeta_1^{t+1} + \zeta_2^{t+1}$$

- 4) Repeat 2)-3) until convergence, and we get $\theta^* = \zeta_1^* + \zeta_2^*$

Information-Geometrical view of Belief Propagation

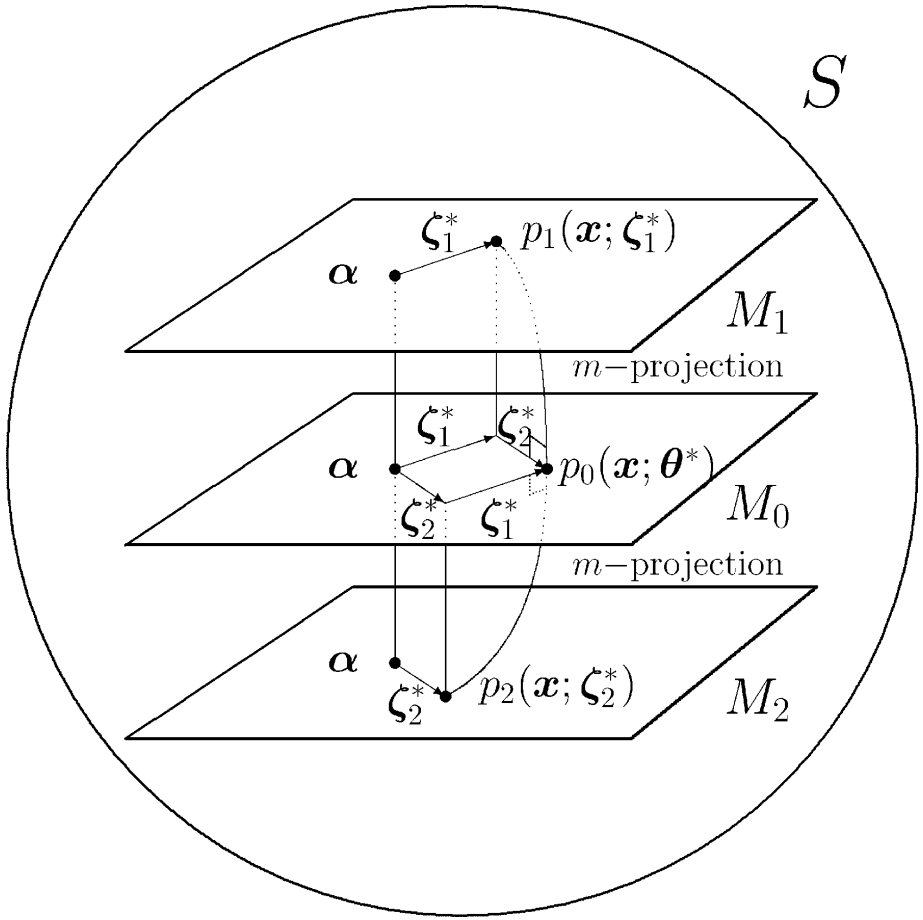


Figure: Information geometrical view of the BP algorithm with L = 2

Information-Geometrical view of Belief Propagation

Analysis of the example with L = 2

The convergence point satisfies the two conditions:

- 1. m-condition: $\theta^* = \prod_{M_0} p_1(\mathbf{x}, \zeta_1^*) = \prod_{M_0} p_2(\mathbf{x}, \zeta_2^*)$
- 2. e-condition: $\theta^* = \zeta_1^* + \zeta_2^* = \xi_1^* + \xi_2^*$

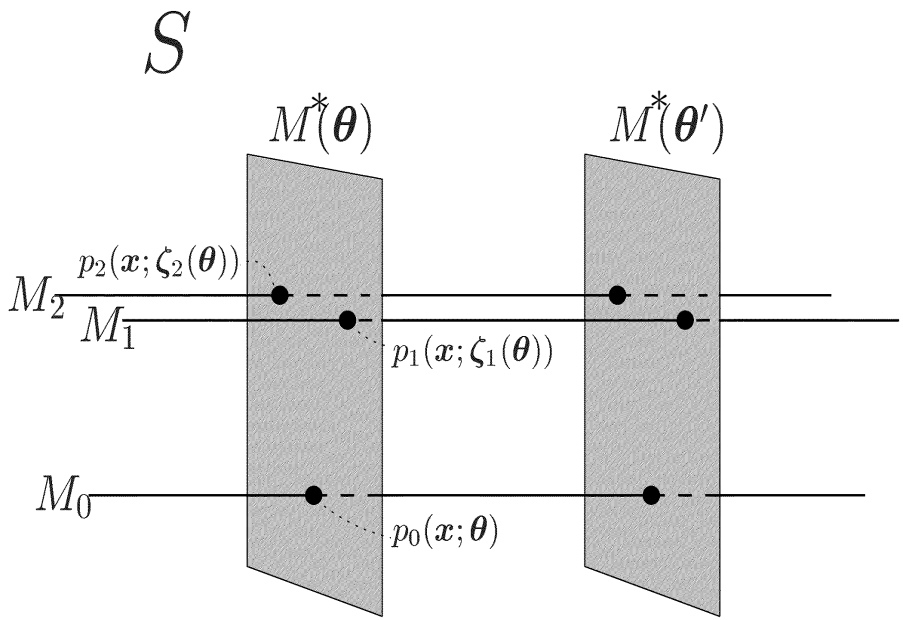
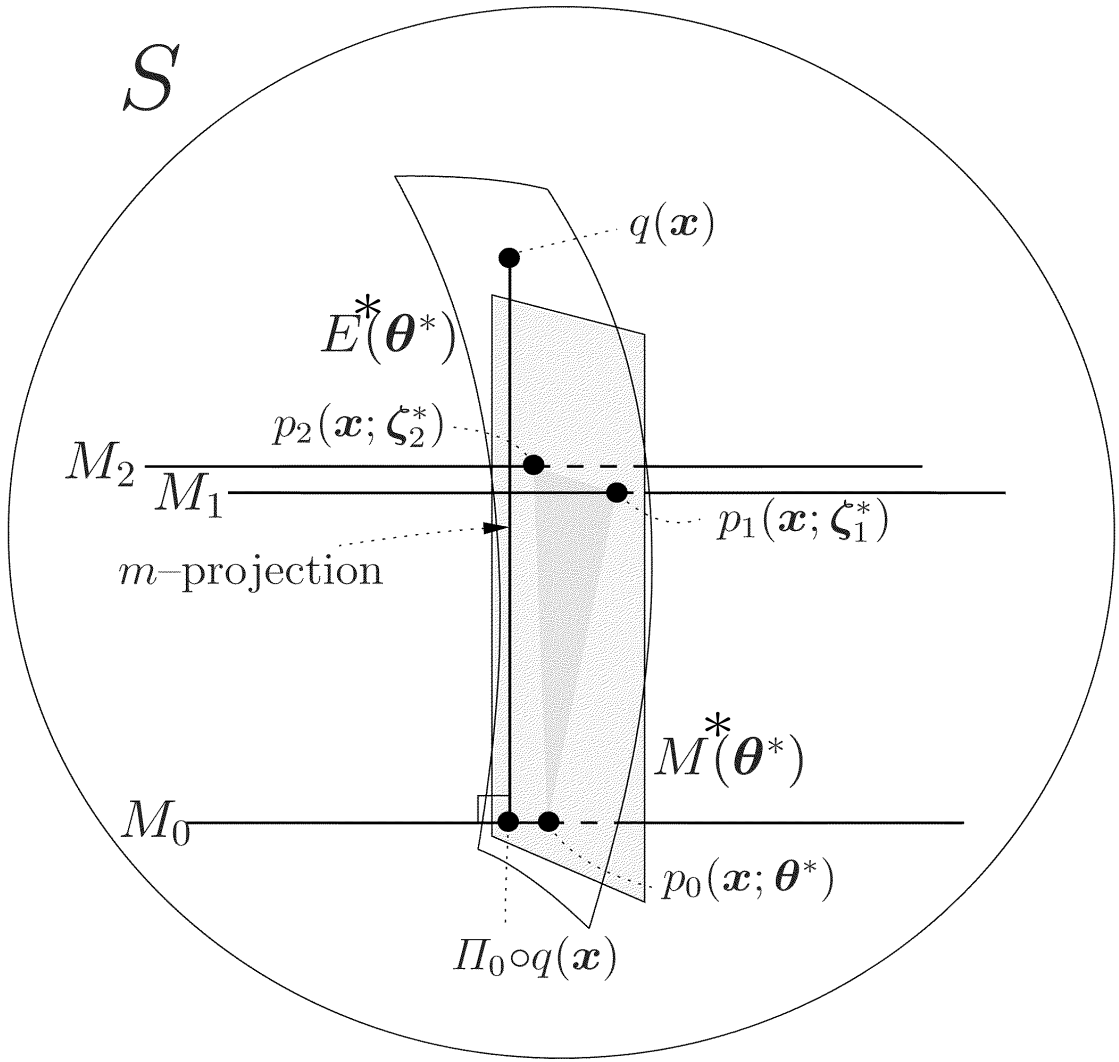


Figure: Equimarinal submanifold M^*

Information-Geometrical view of Belief Propagation



Information-Geometrical view of CCCP-Bethe

Information Geometric Convex Concave Computational Procedure(CCCP-Bethe)

Inner loop: Given $\theta^{\mathbf{t}}$, calculate $\{\zeta_{\mathbf{r}}^{\mathbf{t+1}}\}$ by solving

$$\prod_{M_0} p_r(\mathbf{x}, \zeta_{\mathbf{r}}^{\mathbf{t+1}}) = L\theta^{\mathbf{t}} - \sum_{r=1}^{L} \zeta_{\mathbf{r}}^{\mathbf{t+1}} = p_0(\mathbf{x}, \theta^{\mathbf{t}}), \quad r = 1, \cdots, L$$

outer loop: Given a set of $\{\zeta_{\mathbf{r}}^{\mathbf{t+1}}\}$ as the result of the inner loop, update

$$\theta^{\mathbf{t+1}} = L\theta^{\mathbf{t}} - \sum_{r=1}^{L} \zeta_{\mathbf{r}}^{\mathbf{t+1}}, \quad r = 1, \cdots, L$$

Notes: CCCP enforces the m-condition at each iteration, the e-condition is satisfied only at the convergent point, which means we search for new $\theta^{\mathbf{t+1}}$ until the e-condition is satisfied.

A New e-Constraint Algorithm

Define a cost function

$$\mathcal{F}_e(\{\zeta_r\}) = \sum_r ||\eta_0(\theta) - \eta_r(\zeta_r)||^2$$

under the e-constraint $\theta = \sum_r \zeta_r/(L-1)$. We minimize $\mathcal{F}_e$ by the gradient descent algorithm. The gradient is

$$\frac{\partial \mathcal{F}_e}{\partial \zeta_r} = -2I_r(\zeta_r)[\eta_0(\theta) - \eta_r(\zeta_r)] + \frac{2}{L-1} I_0(\theta) \sum_r [\eta_0(\theta) - \eta_r(\zeta_r)]$$

Here, $I_0(\theta)$ and $I_r(\zeta_r)$ are the Fisher information matrices of $p_0(\mathbf{x}, \theta)$ and $p_r(\mathbf{x}, \zeta_r)$, resepctively, which are defined as

$$I_0(\zeta_r) = \partial_\theta \eta_0(\theta) = \partial_\theta^2 \psi_0(\theta), \quad I_r(\zeta_r) = \partial_{\zeta_r} \eta_r(\zeta_r) = \partial_{\zeta_r}^2 \psi_r(\zeta_r),$$
$$r = 1, \ldots, L.$$

A New e-Constraint Algorithm

Under the gradient descent algorithm, $\zeta_{\mathbf{r}}$ and θ are updated as

$$\zeta_r^{t+1} = \zeta_r^t - \delta \frac{\partial \mathcal{F}_e}{\partial \zeta_r^t}$$

$$\boldsymbol{\theta}^{t+1} = \frac{1}{L-1} \sum_r \zeta_r^{t+1}$$

where δ is a small positive learning rate.

A New e-Constraint Algorithm

A New e–Constraint Algorithm

1. Set $t = 0$, $\boldsymbol{\theta}^t = \mathbf{0}$, $\boldsymbol{\zeta}_r^t = \mathbf{0}$, $r = 1, \ldots, L$.
2. Calculate $\eta_0(\boldsymbol{\theta}^t)$, $I_0(\boldsymbol{\theta}^t)$, and $\eta_r(\boldsymbol{\zeta}_r^t)$, $r = 1, \ldots, L$.
3. Let $\boldsymbol{h}_r = \eta_0(\boldsymbol{\theta}^t) - \eta_r(\boldsymbol{\zeta}_r^t)$ and calculate $\eta_r(\boldsymbol{\zeta}_r^t + \alpha \boldsymbol{h}_r)$ for $r = 1, \ldots, L$, where $\alpha > 0$ is small. Then calculate

$$\boldsymbol{g}_r = \frac{\eta_r(\boldsymbol{\zeta}_r^t + \alpha \boldsymbol{h}) - \eta_r(\boldsymbol{\zeta}_r^t)}{\alpha}.$$

4. For $t = 1, 2, \ldots,$ update $\boldsymbol{\zeta}_r^{t+1}$ as follows:

$$\boldsymbol{\zeta}_r^{t+1} = \boldsymbol{\zeta}_r^t - \delta \left[-2\boldsymbol{g}_r + \frac{2}{L-1} I_0(\boldsymbol{\theta}^t) \sum_r \boldsymbol{h}_r \right],$$

$$\boldsymbol{\theta}^{t+1} = \frac{1}{L-1} \sum_r \boldsymbol{\zeta}_r^{t+1}.$$

5. If $\mathcal{F}_e(\{\boldsymbol{\zeta}_r\}) = \sum_r ||\eta_0(\boldsymbol{\theta}) - \eta_r(\boldsymbol{\zeta}_r)||^2 > \epsilon$ (ϵ is a threshold) holds, $t+1 \rightarrow t$ and go to 2.

Thanks!

Thanks!