Information Geometry and Minimum Description Length Networks

K. Sun*, J. Wang†, A. Kalousis‡*, S. Marchand-Maillet*

* University of Geneva
† Expedia
‡ University of Applied Sciences Western Switzerland
A Statistical Manifold $\mathcal{S}$

$\mathcal{S} = \{ \theta : p(x | \theta) \text{ has certain structures} \}$

A point $\theta \in \mathcal{S}$ is a probability distribution.

Learning forms a path $\theta^0 \rightarrow \theta^1 \rightarrow \cdots$

Geometry of $\mathcal{S}$, defined by

- Fisher Information Metric (FIM) [Rao45]
- $\alpha$-connections [Čencov82][Amari00]
- Divergence [Csiszár63][Bregman67][Amari00]
Let $\mathcal{S}$ be an Exponential Family

In an exponential family $\mathcal{S}$ with a base measure $\sigma(x)$,

$$p(x \mid \theta) = \exp \left( \theta^T t(x) - \psi(\theta) \right) \quad (\psi: \text{convex})$$

$\mathcal{S}$ has two coordinate systems

- canonical parameters $\theta$
- expectation parameters $\eta = E_{p(x \mid \theta)}(t(x))$

The coordinate transformation $\theta \leftrightarrow \eta$ is given by [Amari00]

$$\eta = \frac{\partial \psi}{\partial \theta} \quad \text{and} \quad \theta = \frac{\partial \psi^*}{\partial \eta},$$

where $\psi^* = \int p(x \mid \theta) \ln p(x \mid \theta) d\sigma(x)$ is the negative entropy.
Let $\mathcal{S}$ be Gaussian

A multi-variate Gaussian distribution with mean $\mu$ and covariance $\Sigma$ has the form

$$p(x \mid \mu, \Sigma) \propto \exp \left( \mu^T \Sigma^{-1} x - \frac{1}{2} \text{tr} \left( \Sigma^{-1} xx^T \right) \right)$$

- Canonical parameters: $\theta^{(1)} = \Sigma^{-1} \mu$, $\theta^{(2)} = -\frac{1}{2} \Sigma^{-1}$;
- Expectation parameters: $\eta^{(1)} = \mu$, $\eta^{(2)} = \Sigma + \mu \mu^T$.

$G(\cdot \mid \mu, \Sigma)$

$0$  $\mu$

$\Sigma$

$\sqrt{|g|}$ (g: FIM)
Information Divergence

**Geodesic Distance** $\delta(\eta_1, \eta_2)$ is too complex

**Divergence** $D(\eta_1 \parallel \eta_2)$ is a convenient dissimilarity measure

\[
D(\eta_1 \parallel \eta_2) = \psi^*(\eta_1) - \psi^*(\eta_2) - \frac{\partial \psi^*}{\partial \eta} \bigg|_{\eta=\eta_2} (\eta_1 - \eta_2)
\]

\[
= \cdots = \psi^*(\eta_1) - \eta_1^T \theta_2 + \psi(\theta_2)
\]

This Kullback-Leibler (KL) divergence is among a large family of divergence measures [Csiszár63][Bregman67]
A Mixture Model

\[ p(x) = \sum_{i=1}^{m} \alpha_i p(x \mid \theta_i), \quad \sum_{i=1}^{m} \alpha_i = 1, \quad \forall i, \alpha_i \geq 0. \]  

(1)
A Geometric View of Mixture Learning [Amari95]

\[ \{x_i\}_{i=1}^n \quad \text{i.i.d. samples in} \ \mathbb{R}^d \]

\[ \{y_i\}_{i=1}^n \quad \text{corresponding mixture components (discrete hidden variable)} \]

**Data sub-manifold** \( \mathcal{D} = \{p(x, y)\} \)
Spanned by all possible \( p(y_i) \) while fixing \( x_i \)

**Model sub-manifold** \( \mathcal{M} = \{p(x, y)\} \)
Spanned by all possible mixture models on \( S \) with \( m \) components

**Learning:** finding \( \theta^D \in \mathcal{D} \) and \( \theta^M \in \mathcal{M} \) with minimal divergence
proposed method

simple perception  complex perception  observations
Divergence-induced Priors

In $\mathbb{R}^d$ distance $\rightarrow$ probability:

$$p(x | x_0) \propto \exp \left( -\frac{1}{2} \| x - x_0 \|_2^2 \right)$$

In $S$ divergence $\rightarrow$ probability:

$$p(\eta | \eta_0) \propto \exp (-D(\eta \| \eta_0)) \quad \text{(over a compact region on } S)$$

equal-divergence contours

\[ \eta_0 \]

\[ S \]

\[ p(\eta | \eta_0) \]

high

low
Divergence-induced Priors

- $\mathcal{B} = \{\eta_1, \ldots, \eta_m\} \subset S$
- $\alpha = (\alpha_1, \ldots, \alpha_m)$ s.t. $\sum_{i=1}^m \alpha_i = 1$, $\forall i, \alpha_i > 0$

$$p(\eta | \mathcal{B}, \alpha) \propto \sum_{i=1}^m \alpha_i \exp(-D(\eta \parallel \eta_i)),$$

(2)

- $\ln \left( \sum_{i=1}^m \alpha_i \exp(-D(\eta \parallel \eta_i)) \right) \geq \sum_{i=1}^m \alpha_i (-D(\eta \parallel \eta_i))$
- Like a kernel density estimator on $S$

\begin{tikzpicture}
  \node at (0,0) {\textbf{S}};
  \node at (-1.5,0) {$\eta_1$};
  \node at (1.5,0) {$\eta_2$};
  \fill[black] (-1,0) circle (0.5);
  \fill[black] (1,0) circle (0.5);
  \fill[gray] (-2,0) circle (1);
  \fill[gray] (2,0) circle (1);
  \fill[white] (-3,0) circle (2);
  \fill[white] (3,0) circle (2);
  \shade[shading=ball, ball color=black!20] (-1,0) circle (0.5);
  \shade[shading=ball, ball color=black!40] (1,0) circle (0.5);
  \shade[shading=ball, ball color=gray!20] (-2,0) circle (1);
  \shade[shading=ball, ball color=gray!40] (2,0) circle (1);
  \shade[shading=ball, ball color=white!20] (-3,0) circle (2);
  \shade[shading=ball, ball color=white!40] (3,0) circle (2);

  \fill[black] (-1.5,0) circle (0.05);
  \fill[black] (1.5,0) circle (0.05);
  \fill[gray] (-2.5,0) circle (0.05);
  \fill[gray] (2.5,0) circle (0.05);
  \fill[white] (-3.5,0) circle (0.05);
  \fill[white] (3.5,0) circle (0.05);

  \node at (0,-2) {$p(\eta | \{\eta_1, \eta_2\}, \{\frac{1}{2}, \frac{1}{2}\})$};
  \node at (-2,-2) {low};
  \node at (2,-2) {high};
\end{tikzpicture}
The Description Length

The code length [Shannon48] of \( \eta \) is

\[
- \ln \left[ p(\eta \mid B, \alpha) \sqrt{|g(\eta)|} \delta^{\dim S} \right] = - \ln \left( \sum_{i=1}^{m} \alpha_i \exp(-D(\eta \parallel \eta_i)) \right)
+ \ln N(B, \alpha) - \frac{1}{2} \ln |g(\eta)| - \dim S \ln \delta
\]

\( (N(B, \alpha): \text{normalizer}; \ g(\eta): \text{FIM}; \ \delta: \text{precision}) \)

- Learning is based on the red term, which is
  - novelty of the new knowledge \( \eta \) w.r.t. \( B \) and \( \alpha \)
  - sensitive to parameter variations

- Other terms are useful in post-learning model assessment
Samples on the Boundary $\partial S$

Given a single observation $x$, $t(x)$ in the $\eta$-coordinates is on $\partial S$

$$t(x_i)$$

$$\sum_{j=1}^{m} \alpha_j \exp \left( -D(t(x)\|\eta_j) \right) \propto \sum_{j=1}^{m} \alpha_j p(x | \eta_j)$$

divergence-induced prior

sample likelihood

learning

$$\min \left[ -\sum_{i=0}^{n} \ln \left( \sum_{j=1}^{m} \alpha_j \exp \left( -D(t(x_i)\|\eta_j) \right) \right) \right]$$
Minimum Description Length [Rissanen78] Network

\[ E = - \sum_l \sum_i \ln \left( \sum_j \alpha_{l+1,j} \exp \left( -D(\eta_{li} \parallel \eta_{l+1,j}) \right) \right) \]
Implementations

**HARDN**

\[ E \leq \sum_l \sum_i \min_j \left( - \ln \alpha_{l+1,j} + D(\eta_{li} \| \eta_{l+1,j}) \right) \]

**SOFTN**

\[ E \leq \sum_l \sum_i \sum_j \beta_{li}^j \left( \ln \frac{\beta_{li}^j}{\alpha_{l+1,j}} + D(\eta_{li} \| \eta_{l+1,j}) \right) \]

where \( \forall l, i, j, \beta_{li}^j \geq 0 \) and \( \sum_{j=1}^{n_l+1} \beta_{li}^j = 1 \)

**Symmetrized Centroid [Nielsen & Nock09]**

\[ \min \sum_i w_i^L D(\eta_i^L \| \eta) + \sum_j w_j^R D(\eta \| \eta_j^R) \]

w. r. t. some given \( \{(\eta_i^L, w_i^L)\}, \{(\eta_j^R, w_j^R)\} \subset S \times \mathbb{R}^+ \).

**Strategy:** Natural gradient [Amari98] descent
(see paper for details)
Testing Error on Toy Datasets

(small training size)

(large training size)

4 v.s. 3 (6 v.s. 5) shows the effectiveness of letting $\mathcal{L}_2$ regularize $\mathcal{L}_1$
Testing Error on hand-written “1”s

The size of each layer is a hyper-parameter

Effective regularization means replying less on model selection
Theorem 1

If the truth is a finite mixture model w. r. t. \( \{ \eta_i^t \} \), then as the sample size \( n \to \infty \), \( \mathcal{L}^*_1 \) in an optimal MDL network is exactly \( \{ \eta_i^t \} \).

\( \partial S \)

\( \mathcal{L}_0 \) (fixed)

\( \mathcal{L}_1 \)

\( \mathcal{L}_2 \)

\( \mathcal{S} \)

(intuitively, the cost between \( \mathcal{L}_0 \) and \( \mathcal{L}_1 \) will dominate)
The Gain

**Theorem 2**

∀η₁, η₂ ∈ S, η₁ ≠ η₂, then ∃η ∈ S, s.t.

\[ D(\eta_1 \parallel \eta) + D(\eta \parallel \eta_2) \leq D(\eta_1 \parallel \eta_2) - \max\{D(\eta_{lc} \parallel \eta_1) + D(\eta_1 \parallel \eta_{lc}), \]
\[ D(\eta_{rc} \parallel \eta_2) + D(\eta_2 \parallel \eta_{rc})\}, \]

where \( \theta_{lc} = (\theta_1 + \theta_2)/2 \), and \( \eta_{rc} = (\eta_1 + \eta_2)/2 \).

Surprise the “path” becomes shorter after taking an intermediate stop.
v.s. Bayesian Mixture Models

- No heavy integration involved
- Geometric centroids instead of Bayes’ rule

v.s. Neural Networks

- Non-linearity is introduced by symmetrized centroid
- layers: +regularization instead of +flexibility

v.s. Hierarchical Clustering

- The whole network is learned at once
information geometric compactness of a learning network

\[ S \text{ Gaussian } \rightarrow \text{ Bernoulli, categorical, } \cdots \]

\[ D(\cdot \| \cdot) \] other divergence measures

Network Structure sparsity, dropout, \( \cdots \)
Thank You

Codes available at https://git.unige.ch/gitweb/marchand/mdlnetworks

☆ Olivier Schwander helped proof-read the slides
☆ Tom SF Hanines helped with the Dirichlet Process codes

Q & A