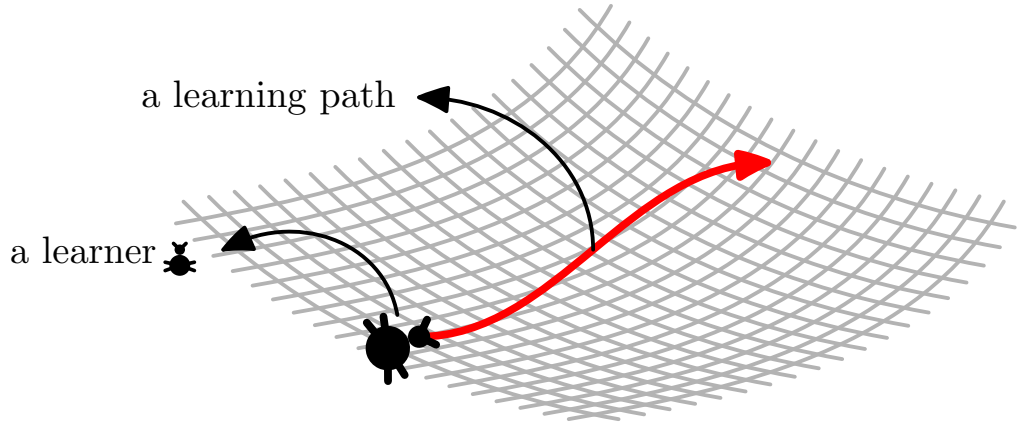
# Information Geometry and Minimum Description Length Networks

K. Sun[⋆], J. Wang[†], A. Kalousis[‡⋆], S. Marchand-Maillet[⋆]

[⋆] University of Geneva
[†] Expedia
[‡] University of Applied Sciences Western Switzerland

# A Statistical Manifold $\mathcal{S}$




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

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

**Learning** forms a path $\boldsymbol{\theta}^0 \to \boldsymbol{\theta}^1 \to \cdots$

**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 $\sigma(\boldsymbol{x})$,

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

$\mathcal{S}$ has two coordinate systems

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

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

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

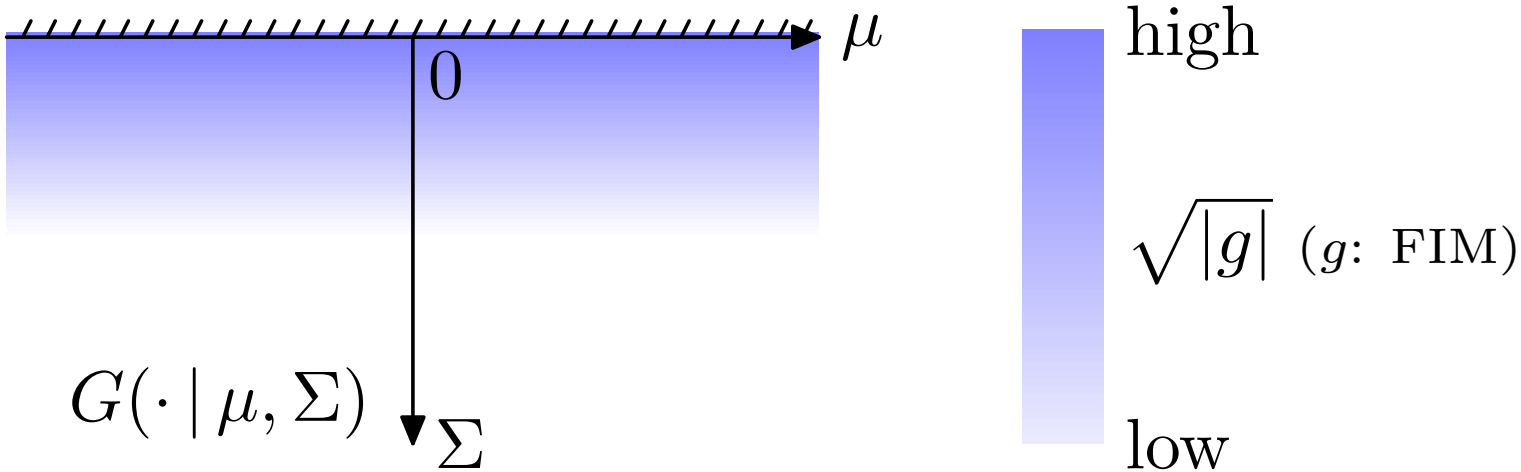
where $\psi^\star = \int p(\boldsymbol{x} \mid \boldsymbol{\theta}) \ln p(\boldsymbol{x} \mid \boldsymbol{\theta}) d\sigma(\boldsymbol{x})$ is the negative entropy.

# Let $\mathcal{S}$ be Gaussian

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

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

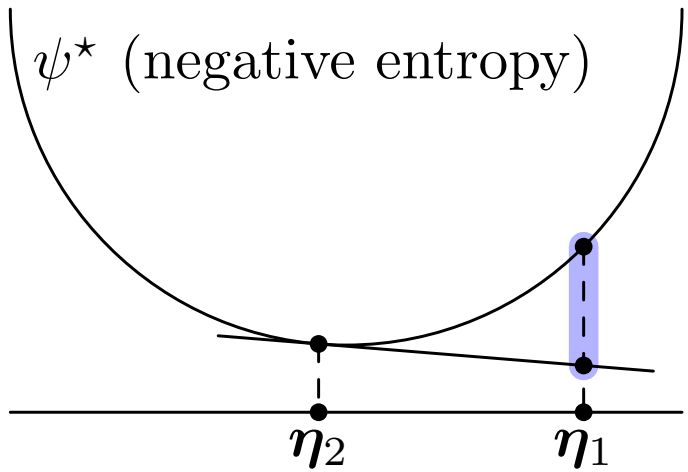
- Canonical parameters: $\boldsymbol{\theta}^{(1)} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, $\boldsymbol{\theta}^{(2)} = -\frac{1}{2}\boldsymbol{\Sigma}^{-1}$;
- Expectation parameters: $\boldsymbol{\eta}^{(1)} = \boldsymbol{\mu}$, $\boldsymbol{\eta}^{(2)} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T$.



# Information Divergence

**Geodesic Distance** $\delta(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ is too complex

**Divergence** $D(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_2)$ is a convenient dissimilarity measure



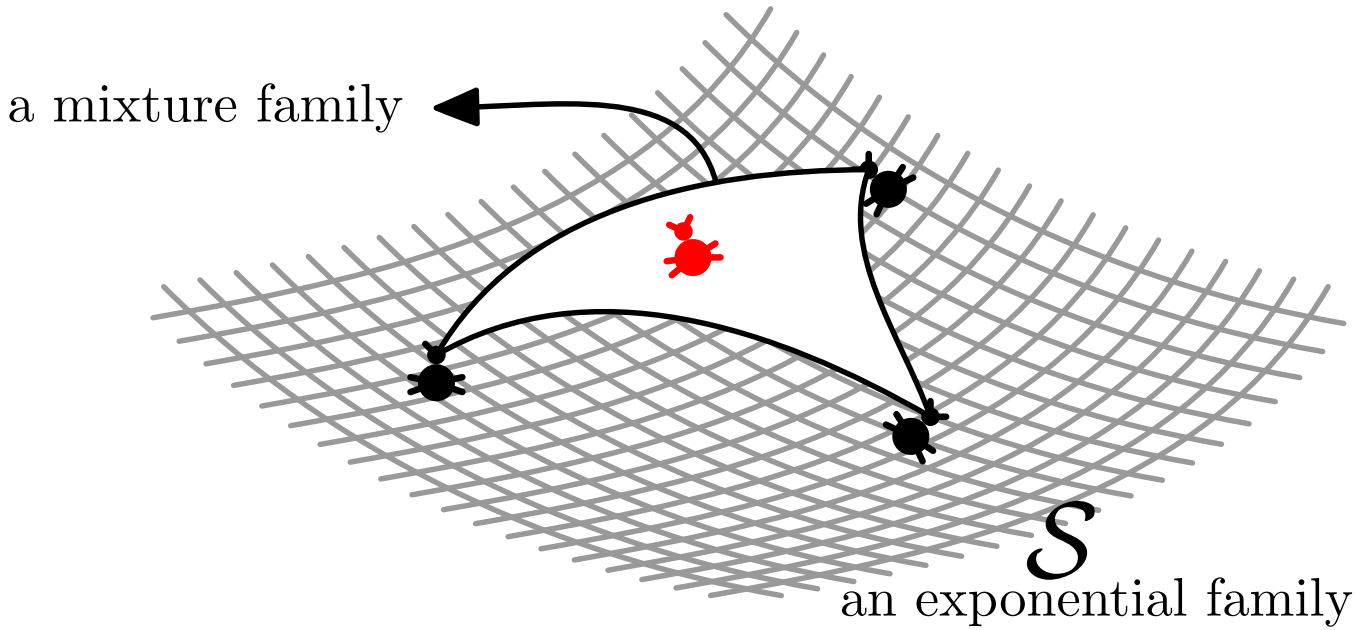

$$D(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_2) = \psi^\star(\boldsymbol{\eta}_1) - \psi^\star(\boldsymbol{\eta}_2) - \left.\frac{\partial \psi^\star}{\partial \boldsymbol{\eta}}\right|_{\boldsymbol{\eta}=\boldsymbol{\eta}_2} (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)$$

$$= \cdots = \psi^\star(\boldsymbol{\eta}_1) - \boldsymbol{\eta}_1^T \boldsymbol{\theta}_2 + \psi(\boldsymbol{\theta}_2)$$

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

# A Mixture Model

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



# A Geometric View of Mixture Learning [Amari95]

$\{x_i\}_{i=1}^n$ i.i.d. samples in $\Re^d$

$\{y_i\}_{i=1}^n$ corresponding mixture components (discrete hidden variable)

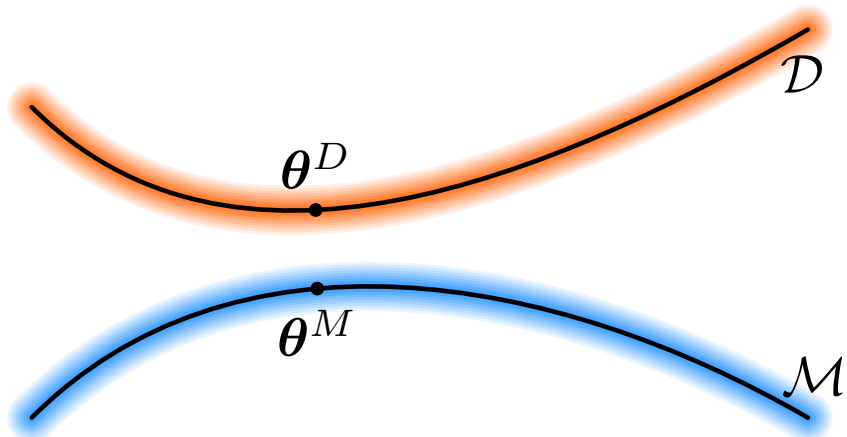
**Data sub-manifold $\mathcal{D} = \{p(x, y)\}$**

Spanned by all possible $p(y_i)$ while fixing $x_i$

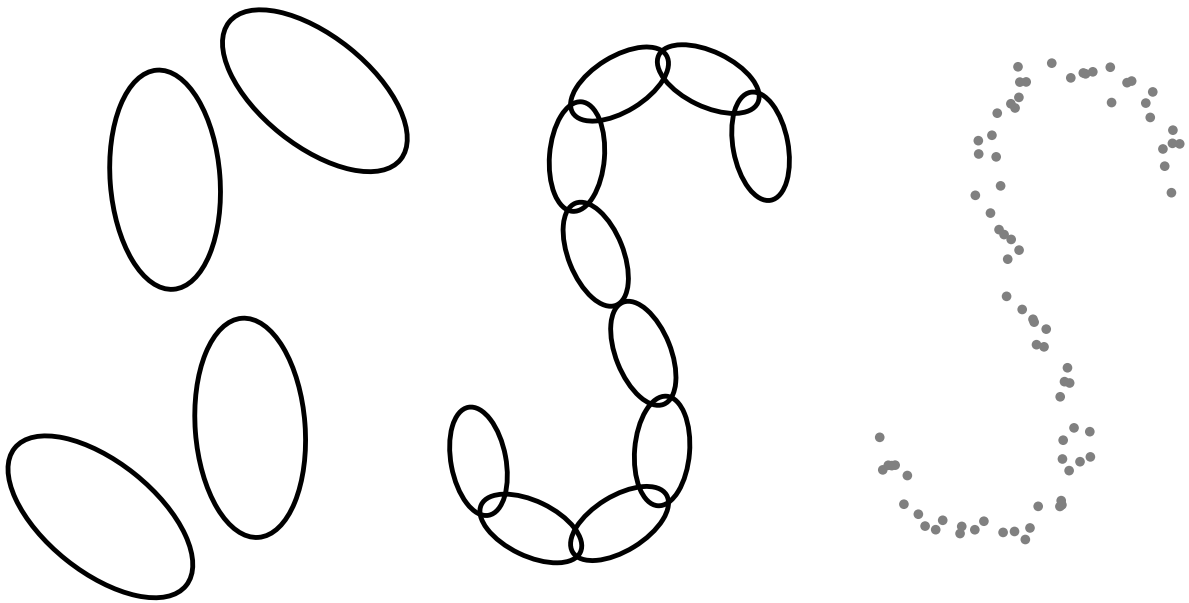
**Model sub-manifold $\mathcal{M} = \{p(x, y)\}$**

Spanned by all possible mixture models on $\mathcal{S}$ 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

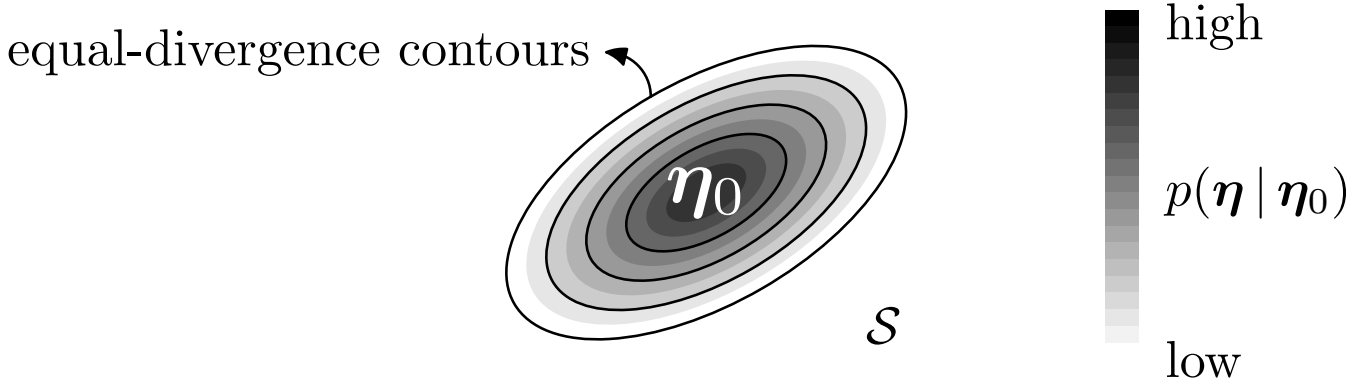
# Divergence-induced Priors

In $\Re^d$ distance $\rightarrow$ probability:

$$p(\boldsymbol{x} \mid \boldsymbol{x}_0) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2\right)$$

In $\mathcal{S}$ divergence $\rightarrow$ probability:

$$p(\boldsymbol{\eta} \mid \boldsymbol{\eta}_0) \propto \exp\left(-D(\boldsymbol{\eta} \,\|\, \boldsymbol{\eta}_0)\right) \quad \text{(over a compact region on } \mathcal{S})$$

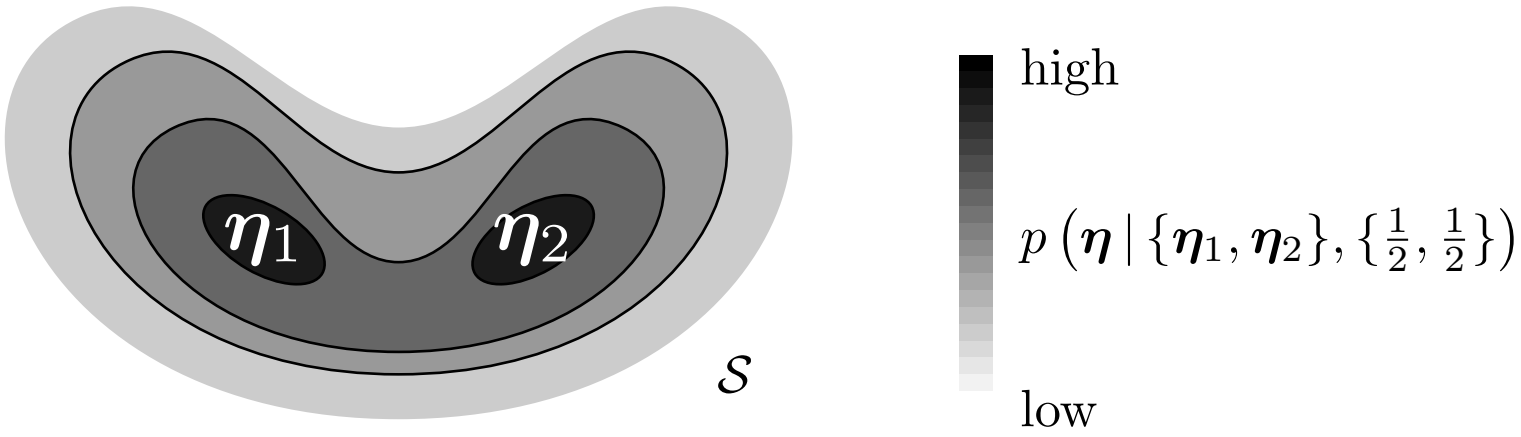


# Divergence-induced Priors

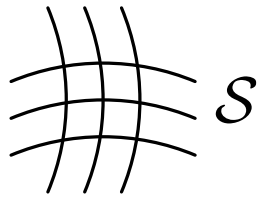
- $\mathcal{B} = \{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m\} \subset \mathcal{S}$
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ s.t. $\sum_{i=1}^{m} \alpha_i = 1$, $\forall i, \alpha_i > 0$

$$p(\boldsymbol{\eta} \mid \mathcal{B}, \boldsymbol{\alpha}) \propto \sum_{i=1}^{m} \alpha_i \exp\left(-D(\boldsymbol{\eta} \,\|\, \boldsymbol{\eta}_i)\right), \tag{2}$$

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



# The Description Length




The code length [Shannon48] of $\boldsymbol{\eta}$ is

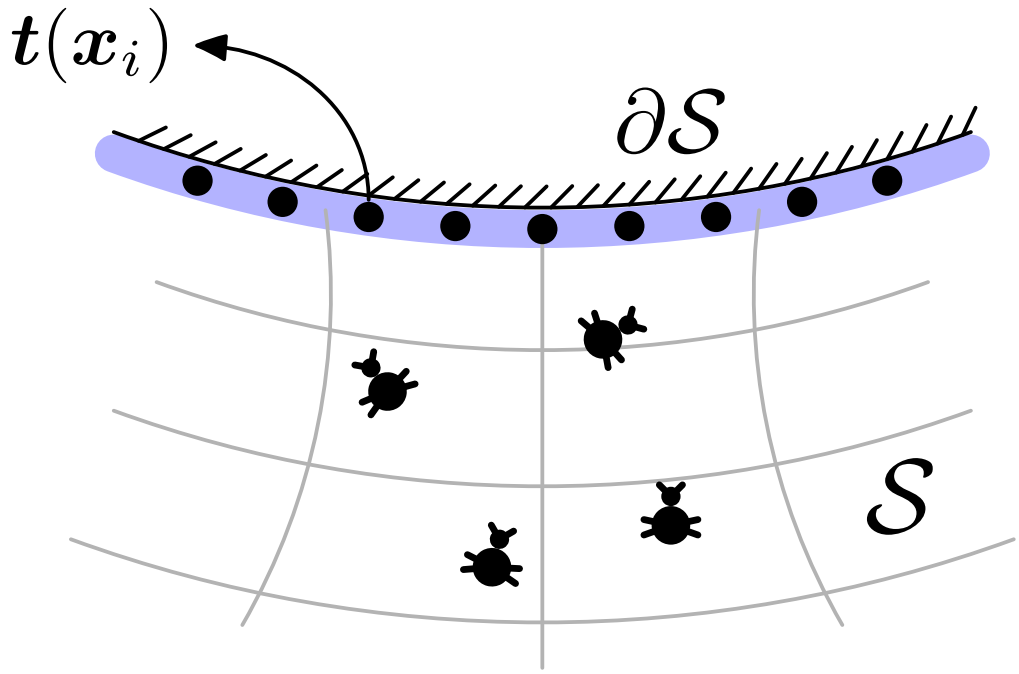
$$-\ln\left[p(\boldsymbol{\eta}\,|\,\mathcal{B},\boldsymbol{\alpha})\sqrt{|g(\boldsymbol{\eta})|}\delta^{\dim\mathcal{S}}\right] = -\ln\left(\sum_{i=1}^{m}\alpha_i \exp\left(-D(\boldsymbol{\eta}\,\|\,\boldsymbol{\eta}_i)\right)\right) + \ln N(\mathcal{B},\boldsymbol{\alpha}) - \frac{1}{2}\ln|g(\boldsymbol{\eta})|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

$\left(\ N(\mathcal{B},\boldsymbol{\alpha})$: normalizer; $g(\boldsymbol{\eta})$: FIM; $\delta$: precision $\right)$

- Learning is based on the red term, which is
  - novelty of the new knowledge $\boldsymbol{\eta}$ w.r.t. $\mathcal{B}$ and $\boldsymbol{\alpha}$
  - sensitive to parameter variations
- Other terms are useful in post-learning model assessment

# Samples on the Boundary $\partial\mathcal{S}$

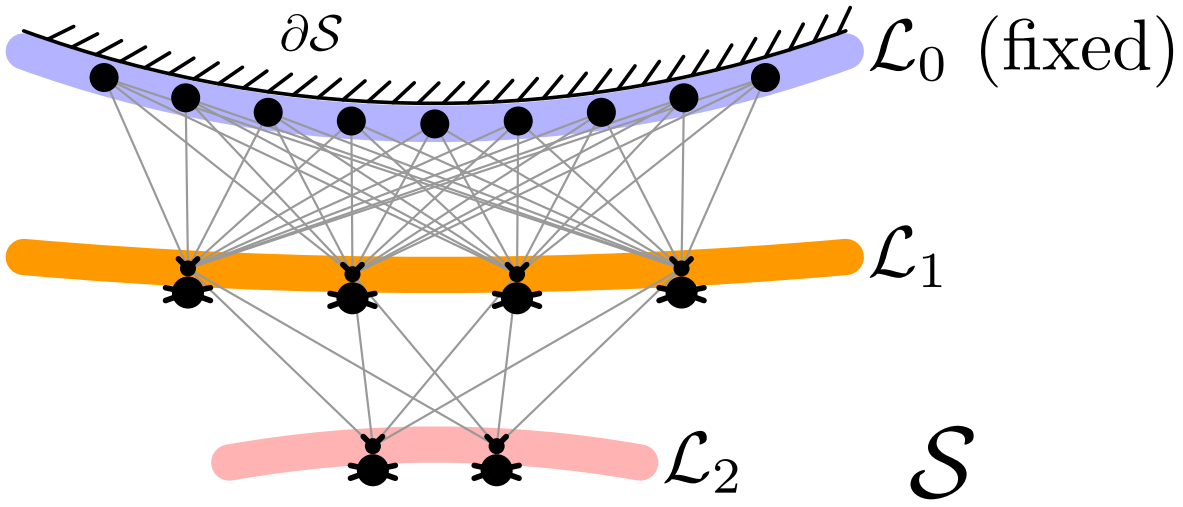
Given a single observation $\boldsymbol{x}$, $\boldsymbol{t}(\boldsymbol{x})$ in the $\eta$-coordinates is on $\partial\mathcal{S}$




$$\underbrace{\sum_{j=1}^{m} \alpha_j \exp\left(-D(\boldsymbol{t}(\boldsymbol{x}) \| \boldsymbol{\eta}_j)\right)}_{\text{divergence-induced prior}} \quad \propto \quad \underbrace{\sum_{j=1}^{m} \alpha_j p(\boldsymbol{x} \mid \boldsymbol{\eta}_j)}_{\text{sample likelihood}}$$

learning $\min \left[ -\sum_{i=0}^{n} \ln \left( \sum_{j=1}^{m} \alpha_j \exp\left(-D(\boldsymbol{t}(\boldsymbol{x}_i) \| \boldsymbol{\eta}_j)\right) \right) \right]$

# Minimum Description Length [Rissanen78] Network




$\mathcal{N}$ a network of mixture components $\{\boldsymbol{\eta}_{li}\} \subset \mathcal{S}$

**layer** $\mathcal{L}_0$ input samples $\{\boldsymbol{t}(\boldsymbol{x}_i)\}$

**layer** $\mathcal{L}_l$ free points (distributions) on $\mathcal{S}$

**shape** $|\mathcal{L}_l| > |\mathcal{L}_{l+1}|, \quad \forall l$

**cost function** $E = -\sum_l \sum_i \ln \left( \sum_j \alpha_{l+1,j} \exp \left( -D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j}) \right) \right)$

# Implementations

**HARDN** $E \leq \sum_l \sum_i \min_j \left( -\ln \alpha_{l+1,j} + D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j}) \right)$

**SOFTN** $E \leq \sum_l \sum_i \sum_j \beta_{li}^j \left( \ln \frac{\beta_{li}^j}{\alpha_{l+1,j}} + D(\boldsymbol{\eta}_{li} \,\|\, \boldsymbol{\eta}_{l+1,j}) \right)$

where $\forall l,\ i,\ j,\ \beta_{li}^j \geq 0$ and $\sum_{j=1}^{n_{l+1}} \beta_{li}^j = 1$

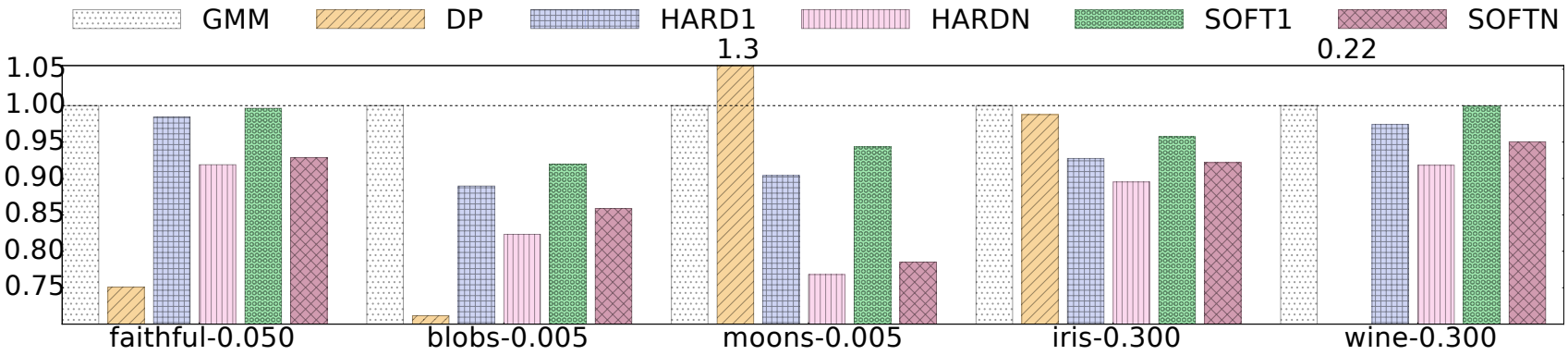
**Symmetrized Centroid [Nielsen & Nock09]**

$$\min \sum_i w_i^L D(\boldsymbol{\eta}_i^L \,\|\, \boldsymbol{\eta}) + \sum_j w_j^R D(\boldsymbol{\eta} \,\|\, \boldsymbol{\eta}_j^R)$$

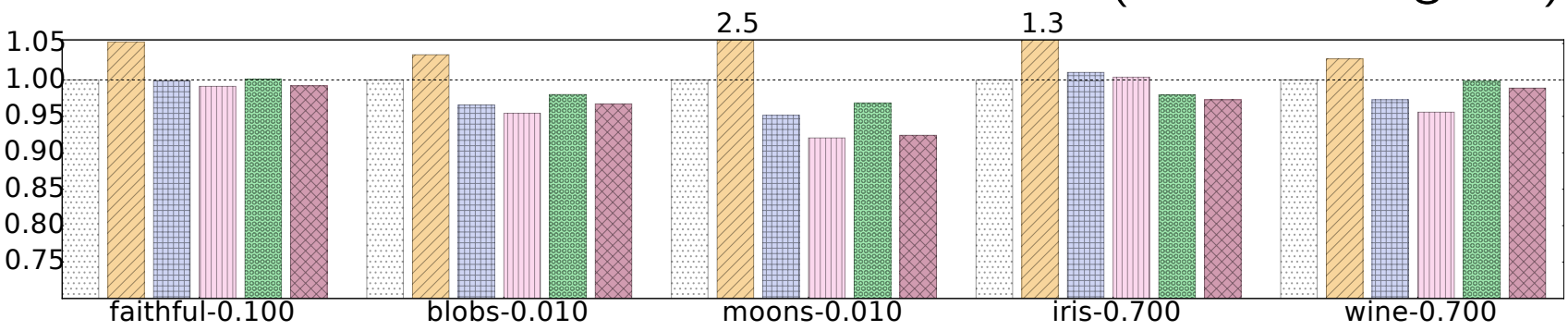
w. r. t. some given $\{(\boldsymbol{\eta}_i^L, w_i^L)\}$, $\{(\boldsymbol{\eta}_j^R, w_j^R)\} \subset \mathcal{S} \times \Re^+$.

**Strategy**: Natural gradient **[Amari98]** descent
(see paper for details)

# Testing Error on Toy Datasets



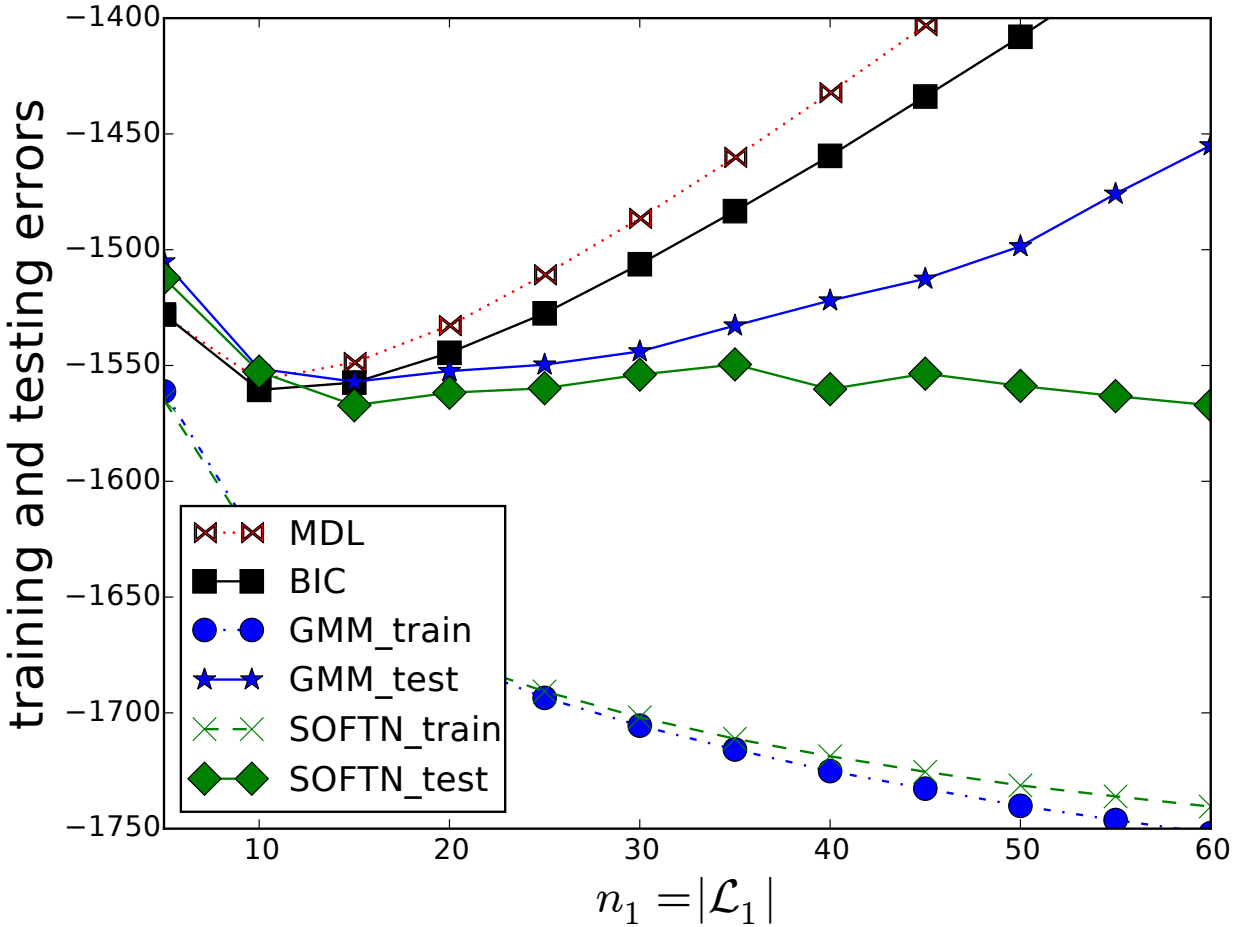

(small training size)




(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



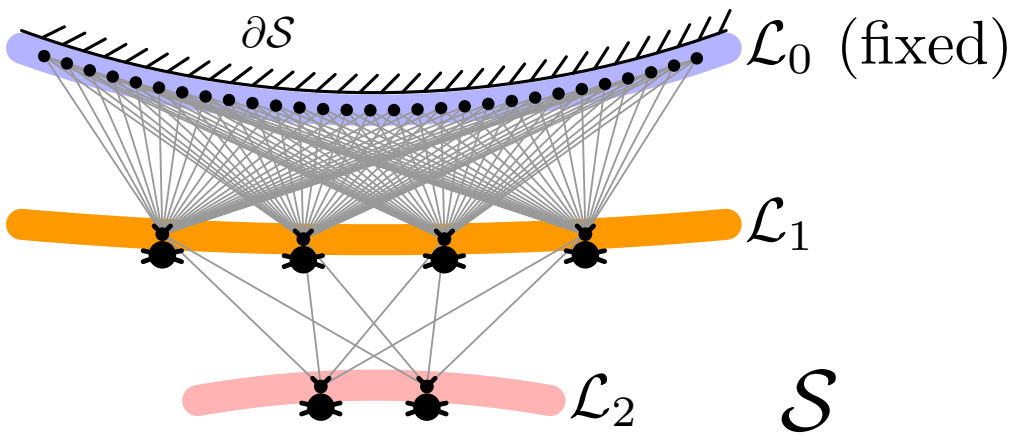

☹ The size of each layer is a hyper-parameter

☺ Effective regularization means replying less on model selection

# Consistency

**Theorem 1**

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




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

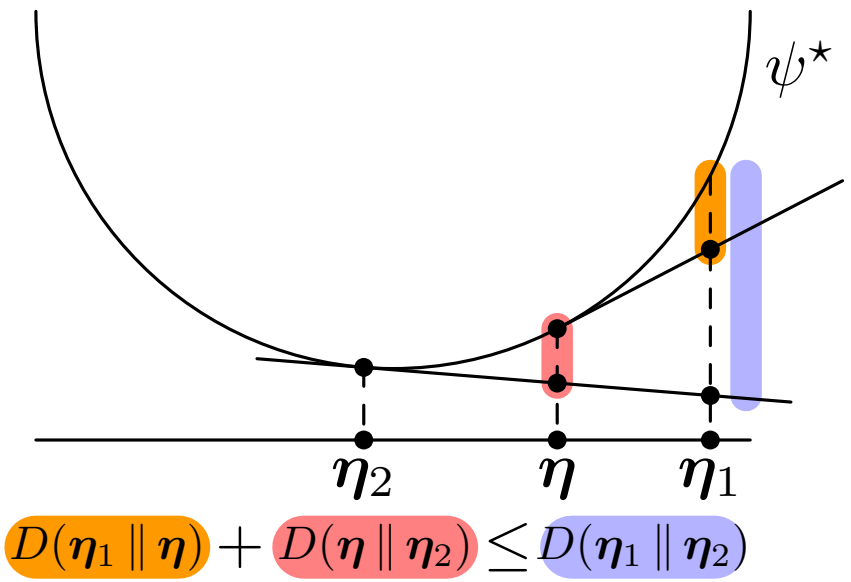
# The Gain

**Theorem 2**

$\forall \boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathcal{S}$, $\boldsymbol{\eta}_1 \neq \boldsymbol{\eta}_2$, *then* $\exists \boldsymbol{\eta} \in \mathcal{S}$, *s.t.*

$$D(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}) + D(\boldsymbol{\eta} \,\|\, \boldsymbol{\eta}_2) \leq D(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_2) - \max\{D(\boldsymbol{\eta}_{lc} \,\|\, \boldsymbol{\eta}_1) + D(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_{lc}), D(\boldsymbol{\eta}_{rc} \,\|\, \boldsymbol{\eta}_2) + D(\boldsymbol{\eta}_2 \,\|\, \boldsymbol{\eta}_{rc})\},$$

*where* $\boldsymbol{\theta}_{lc} = (\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)/2$, *and* $\boldsymbol{\eta}_{rc} = (\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2)/2$.




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

$\mathcal{S}$ Gaussian $\rightarrow$ 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