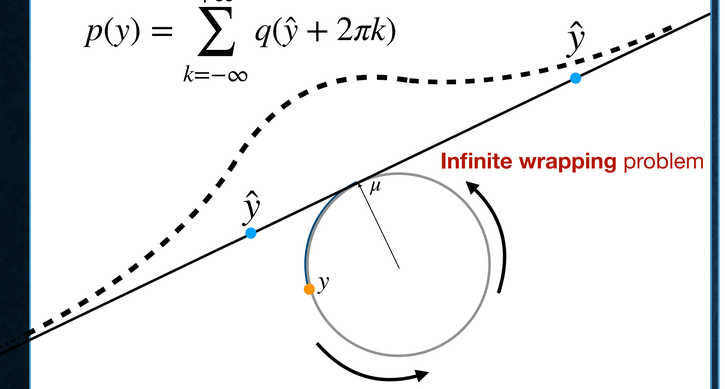
Wrapped β-Gaussians with compact support for exact modelling on manifolds

Sergey Troshin, Vlad Niculae
Language Technology Lab, UvA

Motivation

In general wrapping a distribution over the manifold is **intractable**!

$$p(y) = \sum_{k=-\infty}^{+\infty} q(\hat{y} + 2\pi k)$$



Infinite wrapping problem

Wrapped Gaussian density is intractable for dim=2

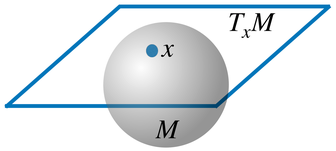
Our solution: all you need is compact support

Background

Differential geometry:

A smooth **manifold** M is a topological space that locally resembles Euclidean space.

Tangent space T_xM at point x is Euclidean

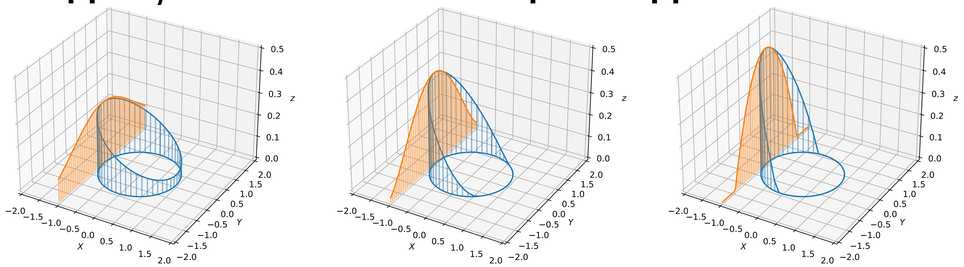


Exp_x, Log_x map between M and T_xM

Learning with sparse support:

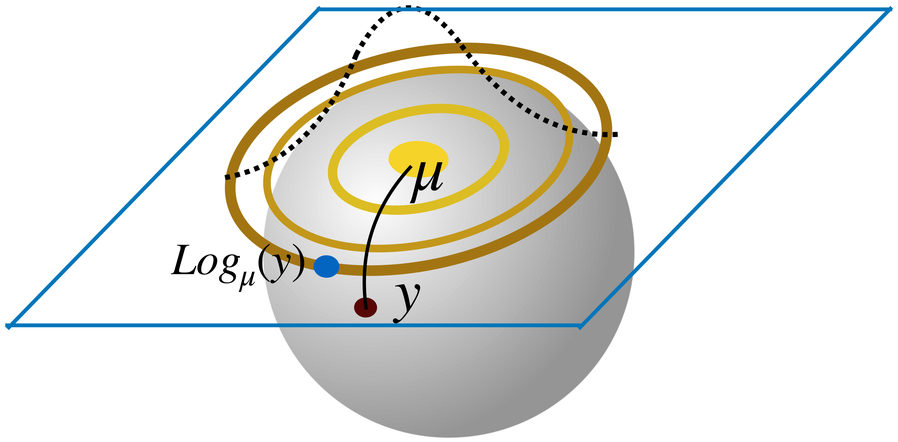
Fenchel Young losses (Martins et al.) generalize **cross-entropy** and **KL** for gradient based learning with **sparse** support

Wrapped β-Gaussians with compact support



β=1 (wrapped gaussian) β=0.5 β=0

$$p(y) = \exp_\beta\left(-\frac{1}{2}Log_\mu(y)^\top \Sigma^{-1} Log_\mu(y) - A_\beta(\Sigma)\right)\left|\frac{\partial Log_\mu(y)}{\partial y}\right|$$



Example: wrapped β-Gaussian for spherical manifold.

We propose:

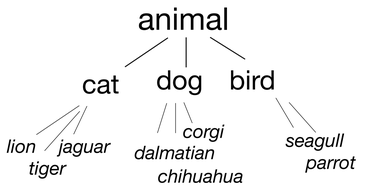
- the **wrapped β-Gaussian**, a distribution family on **Riemannian manifolds**
  - **exact density** assessment
  - **reparametrized sampling** even in high dimensions
  - **anisotropic** scale
- efficient embedded **parametrization** of scale matrices for $\mathbb{S}^n$
- **geometry-aware** Fenchel-Young losses for parameter fitting;

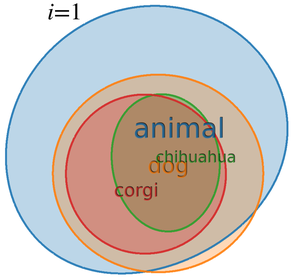
Modelling hierarchies on sphere

Fenchel-Young loss: $L(q:p) \underset{\beta\to 1}{\to} KL(q||p)$

Trading objective:

$$L(w, c_p, c_n) = L(p_w : c_p) - \gamma \log \sum_{i=1}^{k} \exp L(p_w : c_{n_i})$$





Probabilistic word embeddings

Variational Inference on MNIST

$$\log p(x) \geq \mathbb{E}_{q_\theta(z|x)}[\log p_\phi(x|z)] - \mathrm{KL}(q_\theta(z|x)||p(z))$$

z dim	S-VAE	relative $\mathcal{WN}_\beta$-VAE	embedded $\mathcal{WN}_\beta$-VAE
2	-132.50$_{\pm 0.73}$	-130.10$_{\pm 1.65}$	-131.36$_{\pm 0.81}$
5	-108.43$_{\pm 0.09}$	-107.18$_{\pm 0.16}$	-107.16$_{\pm 0.25}$
10	-93.16$_{\pm 0.31}$	-92.60$_{\pm 0.51}$	-92.78$_{\pm 0.25}$
20	-89.02$_{\pm 0.31}$	-97.94$_{\pm 5.84}$	-90.58$_{\pm 0.38}$
40	-90.87$_{\pm 0.34}$	-100.49$_{\pm 4.53}$	-95.62$_{\pm 0.36}$

Variational Inference over rotation matrices

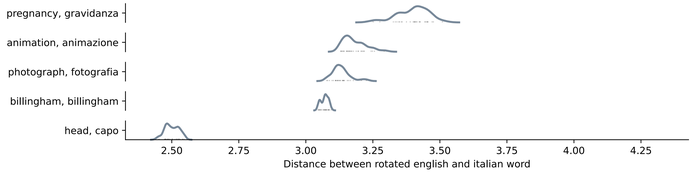
Procrustes problem: $\arg\min\{\sum_i \|u_i - Xv_i\|_2^2 : X \in SO(n)\}$

Rotating **English** space to **Italian:** unsupervised word translation

Approximate posterior over $SO(n)$: $q(X) \sim \mathcal{WN}_\beta(\mu_X, \Sigma_X)$

$$\log p(u \mid v) \geq \mathbb{E}_{q(X)}[\log p(u \mid X, v)] - \mathrm{KL}(q(X) \,\|\, p(X))$$

(prior is uniform under Haar measure)



accuracy: $42.3\% \pm 0.3$ vs *maximum a posteriori* $43.3\,\%$