Overview

- Density estimation with an eye towards matching gradients.
- Kernel exponential family has nice theory, promising applications to relatively simple data so far, but limited by choosing a simple kernel.
- We learn a deep kernel end-to-end with a kind of meta-learning.
- Competitive with deep maximum likelihood models: somewhat worse likelihoods but maybe better “shape”.

Kernel exponential family

- Flexible class of distributions smooth under a given kernel [1, 2, 7].
- Exp. fam. w/ infinite-dimensional parameter \( f \in \mathcal{H} \), an RKHS:
  \[ p_0(x) = \exp(f(x)) \;
  \]
- Normalized density is \( p_0(x) / Z_0 \); \( q \) controls tail behavior.
- Simple \( \mathcal{H} \) gives standard exponent: \( k(x, y) = xy + x^2y^2 \) for Gaussian.
- Richer \( \mathcal{H} \): dense in continuous distributions on compact domain [7].

Density estimation with score matching

- Density estimation: given samples \( \{x_n\}_{n=1}^N \sim p_0 \), want \( p_0 \approx p_0 \).
- Max likelihood: hard to compute \( Z_0 \), often ill-posed in infinite dim.s.
  \[
  J(p_0, p_0) = \frac{1}{2} \sum_{i=1}^D \left\{ \frac{2}{\sigma^2} \log p_0(x) - \left( \frac{1}{2} \sigma^2 \log p_0(x) \right)^2 \right\} + C;
  \]
  doesn’t depend on \( Z_0 \), depends on \( p_0 \) only as expectation.
- Cares about gradient of estimate, unlike max likelihood.
- Get kexpfit with a linear system [7]: good theory, but \( \mathcal{O}(N^2D^2) \).
- Speed up by finding \( f \) in subspace of \( \mathcal{H} \) [8, 9]; \( \mathcal{O}(MD) \) time with \( f(x) = \sum_m \alpha_m k(x, z_m) \); needs \( \partial_{x_m} k(x, z_m) ; \partial_{x_m}^2 k(x, z_m) \).

Need for a learned kernel

- Left: no single bandwidth works well for both peaks.
- Right: a learned kernel scales to the inherent variation of the data.

Learning \( \theta \): kernel, \( z, q \), regularization parameters

- Take \( \nabla_{\theta} \phi \) of this whole process – meta-learning, using closed form \( \hat{\alpha} \).

A class of flexible deep kernels

\[
 k(x, y) = \sum_{i=1}^R \rho_i \phi_\theta(\|x_i - y_i\|)
\]

Results on synthetic problems

- \( \hat{\alpha} \) is a deep network: three FC softplus layers, skip connection from first to last. Guarantees \( Z_0 < \infty \).

Results on real data

- FSSD [4] is a measure of relative model fit based on \( \nabla_x \log \hat{p}(x) \).
- Low \( p \)-value indicates DKEF’s FSSD is confidently better.
- Lines on likelihoods indicate estimates of a bias upper bound.
- DKEF likelihoods somewhat worse than [6], FSSDs somewhat better.

Evaluating unnormalized model likelihoods

- Easy unbiased estimator \( \hat{Z}_0 \) for \( Z_0 \) by importance sampling.
- But by Jensen, \( \log \hat{p}(x) = \log \hat{p}(x) - \log \hat{Z}_0 \) is biased upwards.
- If we propose from \( q \): can upper-bound bias in terms of \( \phi(x) / q(x) \)’s infimum (easy loose bound) and median (can bound w.h.p.).
- Can estimate that upper bound; estimator itself biased upwards.

Behavior on separated mixtures

- Consider \( p_0 = \pi p_1 + (1 - \pi) p_2 \), with disjoint, separated support.
- If model has the same disjoint support, score matching doesn’t care at all about relative weight between the two components.
- If kexpfam kernel has components totally separated, fits as if separately, but scaled \( \lambda \) (smaller components regularized more).
- In extremely simplified case: want ratio \( \pi \), get exp \( \left( \frac{\lambda}{2} \right) \).
- Compare to “bridges” from likelihood models.

Bibliography


This work was supported by the Gatsby Charitable Foundation.