Learning Deep Kernels for Exponential Family Densities

Li Kevin Wenliang   Dougal J. Sutherland   Heiko Strathmann   Arthur Gretton

Gatsby Computational Neuroscience Unit, University College London

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. with infinite-dimensional parameter $f \in \mathcal{H}$, an RKHS

$$\tilde{p}_\theta(\mathbf{x}) = \exp\left(f(\mathbf{x})\right) q(\mathbf{x}) = \exp\left(\langle f, k(\mathbf{x}, \cdot)\rangle_{\mathcal{H}}\right) q(\mathbf{x}).$$

- Normalized density is $p_\theta(\mathbf{x}) = \tilde{p}_\theta(\mathbf{x}) / Z_\theta$; q controls tail behavior.
- Simple $\mathcal{H}$ give standard expfams: $k(x, y) = xy + x^2 y^2$ for Gaussian.
- Richer $\mathcal{H}$: dense in continuous distributions on compact domain [7].

Density estimation with score matching

- Density estimation: given samples $\{\mathbf{x}_n\}_{n=1}^N \sim p_0$, want $p_\theta \approx p_0$.
- Max likelihood: hard to compute Z_θ, often ill-posed in infinite dim.s.
- Alternative: minimize Fisher divergence, under mild assumptions [3]

$$J(p_\theta \| p_0) = \frac{1}{2} \int p_0(\mathbf{x}) \left\| \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_0(\mathbf{x}) \right\|^2 \mathrm{d}\mathbf{x}$$

$$= \int p_0(\mathbf{x}) \sum_{d=1}^{D} \left[\frac{\partial^2}{\partial x_d^2} \log p_\theta(\mathbf{x}) + \frac{1}{2} \left(\frac{\partial}{\partial x_d} \log p_\theta(\mathbf{x}) \right)^2 \right] \mathrm{d}\mathbf{x} + C;$$

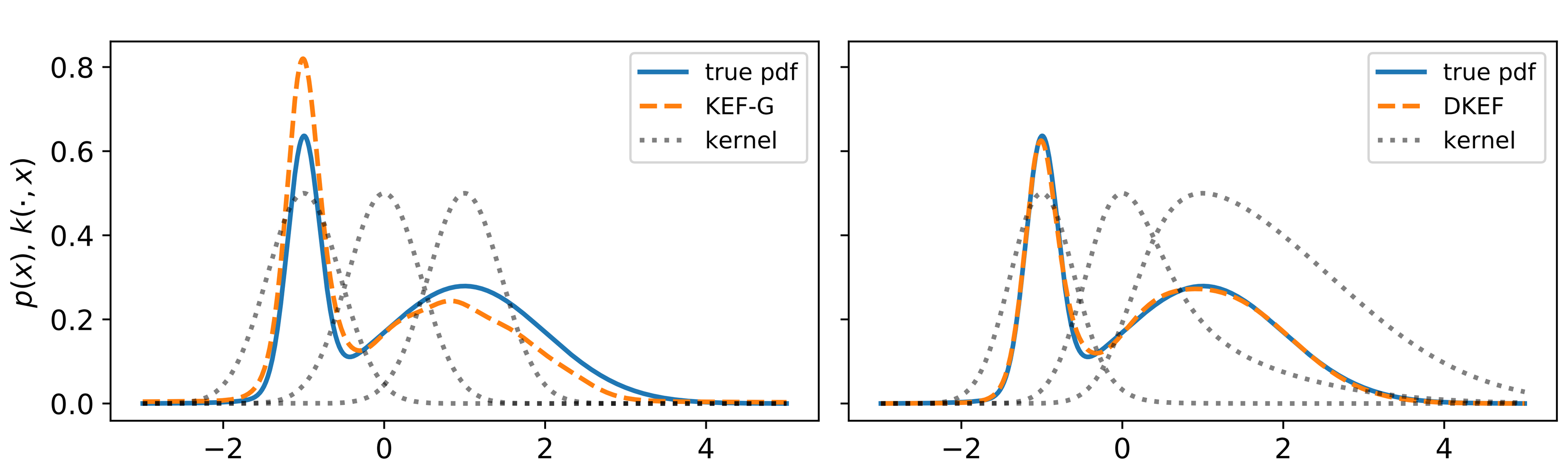
doesn't depend on Z_θ, depends on p_0 only as expectation.

- Cares about gradient of estimate, unlike max likelihood.
- Get kexpfam fit with a linear system [7]; good theory, but $\mathcal{O}(N^3 D^3)$.
- Speed up by finding f in subspace of $\mathcal{H}$ [8, 9]; $\mathcal{O}(M^2 ND)$ time with

$$f(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m k(\mathbf{x}, \mathbf{z}_m);\ \text{needs}\ \frac{\partial}{\partial x_d} k(\mathbf{x}_n, \mathbf{z}_m),\ \frac{\partial^2}{\partial x_d^2} k(\mathbf{x}_n, \mathbf{z}_m).$$

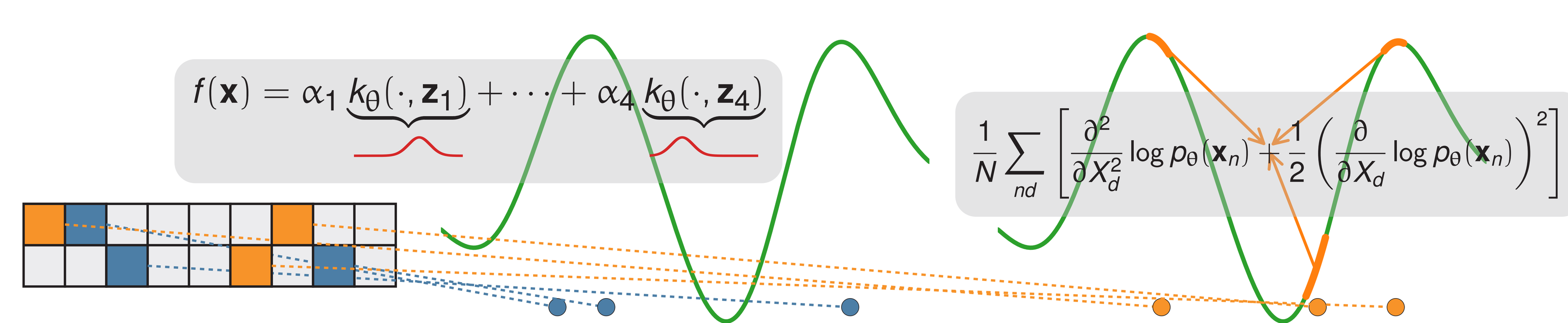
Need for a learned kernel

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



Learning θ: kernel, z, q, regularization parameters

- Take ∇_θ of this *whole process* – meta-learning, using closed form $\hat{\alpha}$:



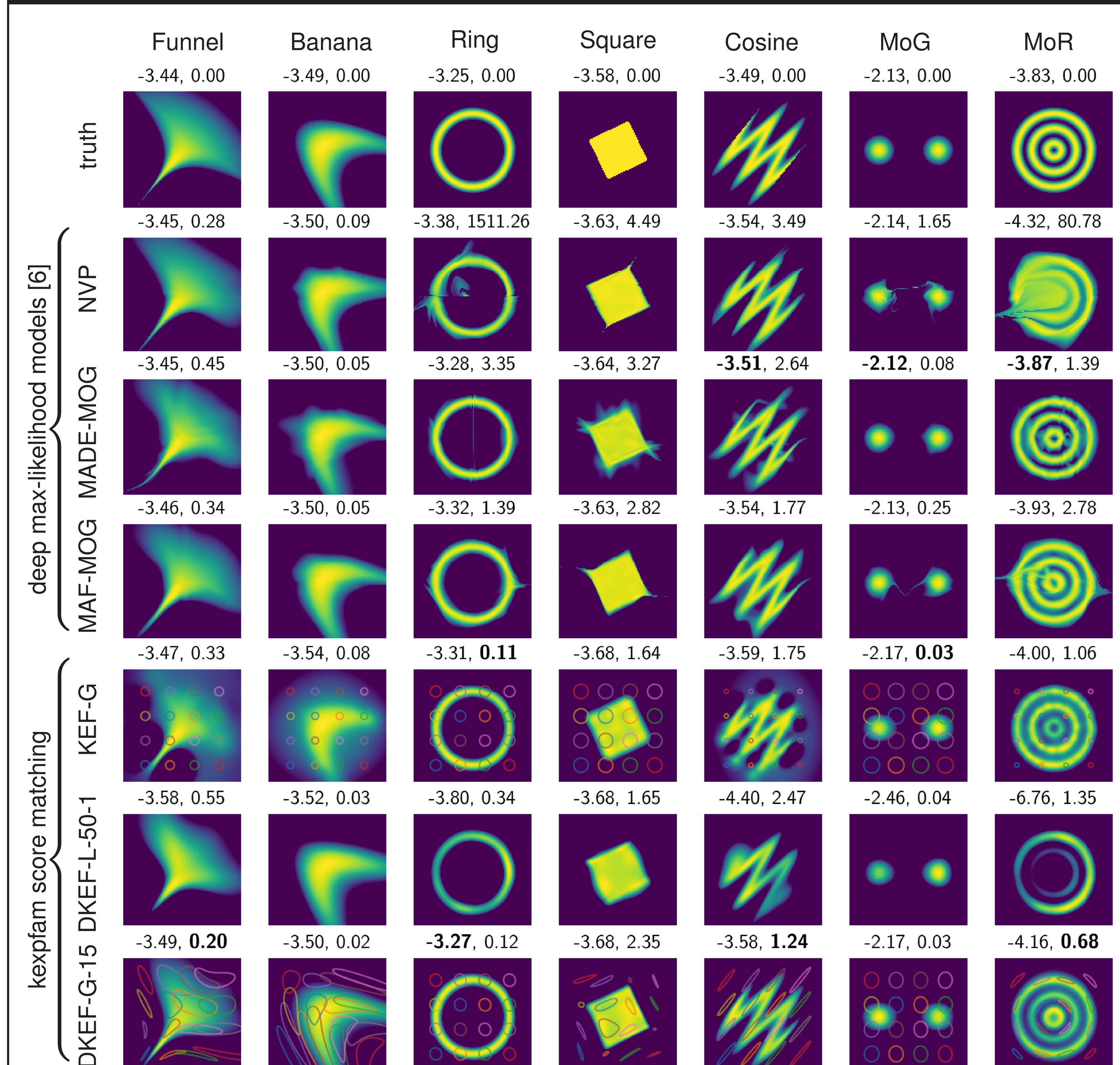
- After fitting kernel, choose reg. λs for bigger batches, then final $\hat{\alpha}$.
- TensorFlow not optimized for $\nabla_\theta\, \texttt{solve}(\{\frac{\partial^2}{\partial x_d^2} k(\mathbf{x}_n, \mathbf{z}_m), \dots\})$; faster after implementing backprop for $\frac{\partial^2}{\partial x_d^2} k(\mathbf{x}_n, \mathbf{z}_m)$ as TensorFlow ops.

A class of flexible deep kernels

$$k(\mathbf{x}, \mathbf{y}) = \sum_{r=1}^{R} \rho_r \exp\left(-\frac{1}{2\sigma_r^2} \left\| \phi_{\mathbf{w}_r}(\mathbf{x}) - \phi_{\mathbf{w}_r}(\mathbf{y}) \right\|^2 \right)$$

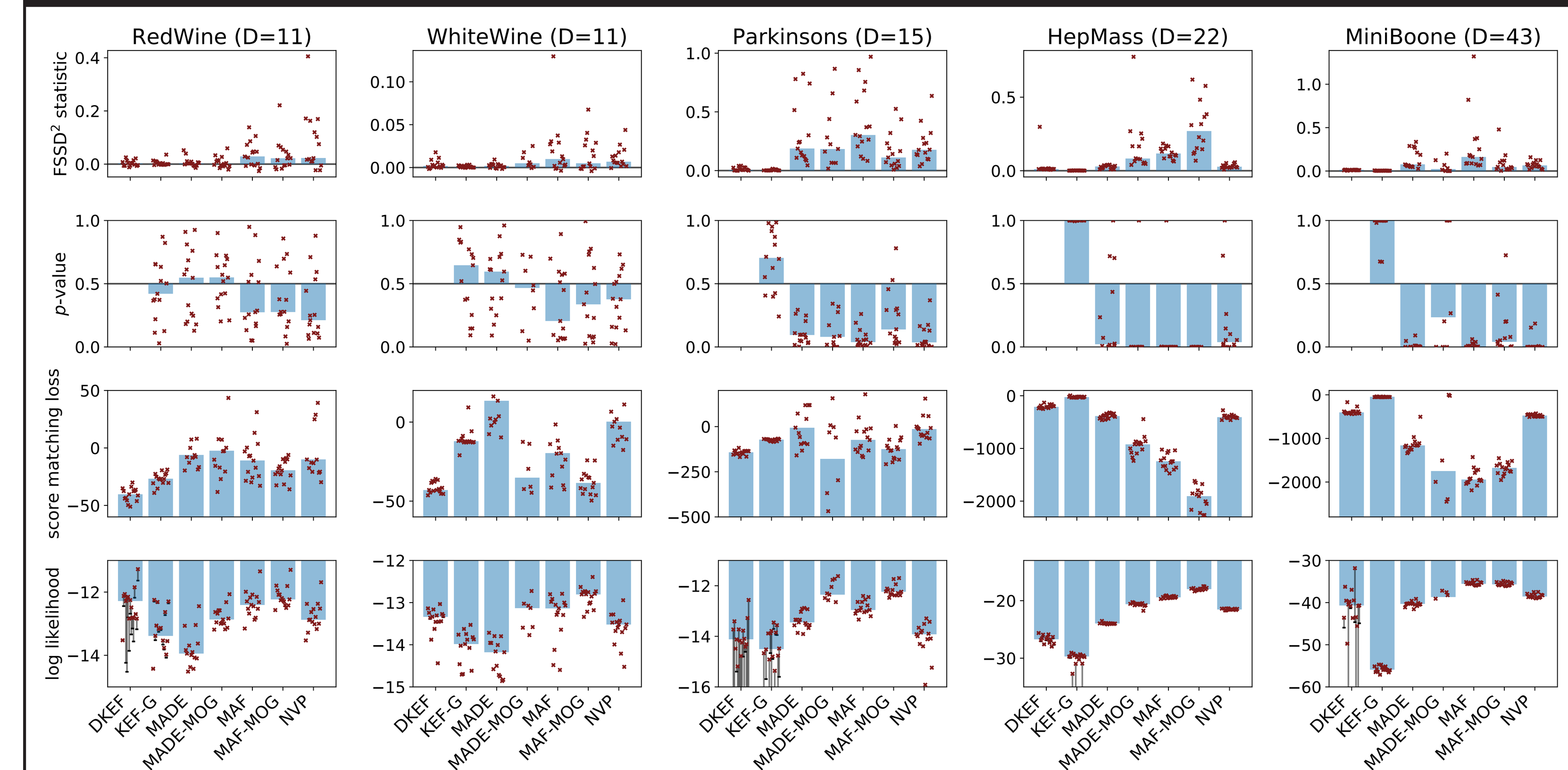
$\phi : \mathbb{R}^D \to \mathbb{R}^{D_r}$ is a deep network: three FC softplus layers, skip connection from first to last. Guarantees $Z_\theta < \infty$.

Results on synthetic problems



DKEF-L-50-1 basically fits a network to the log-density directly [5].

Results on real data



- FSSD [4] is a measure of relative model fit based on $\nabla_{\mathbf{x}} \log \tilde{p}(\mathbf{x})$.
- 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 $\hat{Z}_\theta$ for Z_θ by importance sampling.
- But by Jensen, $\log \hat{p}(\mathbf{x}) = \log \tilde{p}(\mathbf{x}) - \log \hat{Z}_\theta$ is biased upwards.
- If we propose from q: can upper-bound bias in terms of $\tilde{p}(\mathbf{x}) / q(\mathbf{x})$'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 λ_α (smaller components regularized more).
- In extremely simplified case: want ratio $\frac{\pi}{1-\pi}$, get $\exp\left(\frac{D}{2\sigma^2 \lambda_\alpha}\left(\pi - \frac{1}{2}\right)\right)$.
- Compare to "bridges" from likelihood models.

Bibliography

[1] S. Canu and A. J. Smola. "Kernel methods and the exponential family." *Neurocomputing* 69 (7 2006).

[2] K. Fukumizu. "Exponential manifold by reproducing kernel Hilbert spaces." *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, 2009.

[3] A. Hyvärinen. "Estimation of non-normalized statistical models by score matching." *JMLR* 6.Apr (2005).

[4] W. Jitkrittum, H. Kanagawa, Z. Szábo, P. Sangkloy, J. Hays, B. Schölkopf, and A. Gretton. "Informative Features for Model Comparison." *NeurIPS*. 2018.

[5] D. P. Kingma and Y. LeCun. "Regularized estimation of image statistics by score matching." *NIPS*. 2010.

[6] G. Papamakarios, T. Pavlakou, and I. Murray. "Masked Autoregressive Flow for Density Estimation." *NIPS*. 2017.

[7] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar. "Density estimation in infinite dimensional exponential families." *JMLR* 18.1 (2017).

[8] H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabó, and A. Gretton. "Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families." *NIPS*. 2015.

[9] D. J. Sutherland, H. Strathmann, M. Arbel, and A. Gretton. "Efficient and principled score estimation with Nyström kernel exponential families." *AISTATS*. 2018.

This work was supported by the Gatsby Charitable Foundation.

Contact: `wenliang2012@gmail.com`, `dougal@gmail.com`.