Efficient and principled score estimation with Nyström kernel exponential families

Dougal J. Sutherland*, Heiko Strathmann*, Michael Arbel, Arthur Gretton

Gatsby Computational Neuroscince Unit, University College London

Problem: Unnormalized density estimation

- Given samples $\{X_a\}_{a=1}^{n} \overset{iid}{\sim} p_0$, $X_a \in \mathbb{R}^d$
- Want *computationally efficient* estimator p so that $p(x)/Z \approx p_0(x)$
- Don't especially care about Z: often difficult, not needed for finding modes / sampling (with MCMC) / use in approximate HMC / …
- Want to avoid strong (parametric) assumptions about p_0

Exponential families

- Many classic densities on $\mathbb{R}^d$ are of the form:

$$p(x) = \exp\big(\langle \underbrace{\eta}_{\text{natural parameter}}, \underbrace{T(x)}_{\text{sufficient statistic}} \rangle_{\mathbb{R}^s} - \underbrace{A(\eta)}_{\text{log-normalizer}}\big) \underbrace{q_0(x)}_{\text{base measure}}$$

- Gaussian: $T(x) = (x, x^2)$; Gamma: $T(x) = (x, \log x)$
- Density is on $T(x)$, s-dimensional "features"; can we make this richer?

Kernel exponential families [1]

- Use an RKHS $\mathcal{H}$, with kernel $k(x,y) = \langle k_x, k_y \rangle_{\mathcal{H}}$: parameter $\eta = f \in \mathcal{H}$, sufficient statistic $T(x) = k_x$ gives

$$p(x) = \exp\big(f(x) - A(f)\big)\, q_0(x)$$

- Includes standard exponential family: $k(x,y) = T(x) \cdot T(y)$
- But T can be infinite-dimensional, e.g. $k(x,y) = \exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right)$
- Class very rich: dense in anything with smooth log-density, tails like q_0 [3]
- But $A(f)$ is hard to compute: maximum likelihood estimate intractable

Score matching-based estimator [3]

- Score matching approach here: minimize regularized Fisher divergence

$$J_\lambda(f) = \frac{1}{2}\int p_0(x)\, \|\nabla_x \log p_f(x) - \nabla_x \log p_0(x)\|_2^2 \,\mathrm{d}x + \lambda \|f\|_{\mathcal{H}}^2$$

$$= \int p_0(x) \sum_{i=1}^{d} \left[\partial_i^2 f(x) + \frac{1}{2}\left(\partial_i f(x)\right)^2\right] \mathrm{d}x + C(p_0, q_0) + \lambda \|f\|_{\mathcal{H}}^2$$

where we used integration by parts, some mild assumptions

- Estimate integral with simple Monte Carlo
- Representer theorem: best solution $f_{\lambda,n} = \operatorname{argmin}_{f \in \mathcal{H}} \hat{J}_\lambda(f)$ is

$$f_{\lambda,n}(x) = \sum_{a=1}^{n}\sum_{i=1}^{d}\left(\beta_{(a,i)} - \tfrac{1}{\lambda}\partial_i \log q_0(X_a)\right)\partial_i k(X_a, x) - \tfrac{1}{n\lambda}\partial_i^2 k(X_a, x)$$

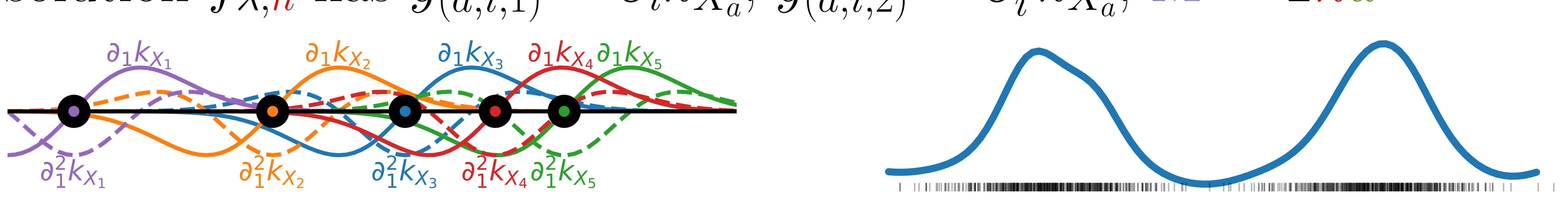
where β is the solution to an $nd \times nd$ linear system: $\mathcal{O}(n^3 d^3)$ time!

Nyström approximation

- Instead of minimizing f over $\mathcal{H}$, minimize over subspace

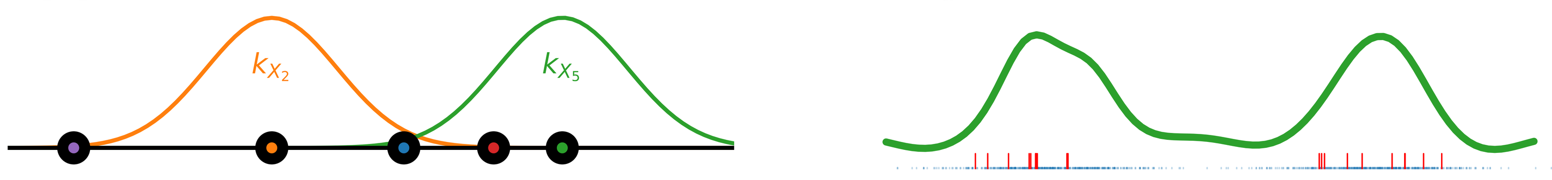
$$\mathcal{H}_Y = \operatorname{span}\{y_b\}_{b=1}^{M} \subset \mathcal{H}$$

- Full solution $f_{\lambda,n}$ has $y_{(a,i,1)} = \partial_i k_{X_a}$, $y_{(a,i,2)} = \partial_i^2 k_{X_a}$; $M = 2nd$



- "Nyström": pick m points at random, $y_{(a,i)} = \partial_i k_{X_a}$; $M = md$

- "lite" [4]: pick m points at random, $y_a = k_{X_a}$; $M = m$



Computing the Nyström approximation

- Minimizer of J_λ in $\mathcal{H}_Y$ is $f_{\lambda,n}^{Y}(x) = \sum_{b=1}^{M} \beta_b\, y_b$,

$$\beta = -\Big(\tfrac{1}{n} \underbrace{B_{XY}^{\mathsf{T}}}_{M \times nd} \underbrace{B_{XY}}_{nd \times M} + \lambda \underbrace{G_{YY}}_{M \times M}\Big)^{\dagger} \underbrace{h_Y}_{M \times 1}$$

$(B_{XY})_{(a,i),j} = \langle \partial_i k_{X_a}, y_j \rangle_{\mathcal{H}}$ $\quad (G_{YY})_{a,b} = \langle y_a, y_b \rangle_{\mathcal{H}}$ $\quad (h_Y)_b = \frac{1}{n}\sum_{a=1}^{n}\sum_{i=1}^{d} \langle \partial_i k_{X_a}, y_b \rangle_{\mathcal{H}}\, \partial_i \log q_0(X_a) + \langle \partial_i^2 k_{X_a}, y_b \rangle_{\mathcal{H}}$

- "Nyström": $\mathcal{O}(nm^2d^3)$ time; "lite": $\mathcal{O}(nm^2d)$ time

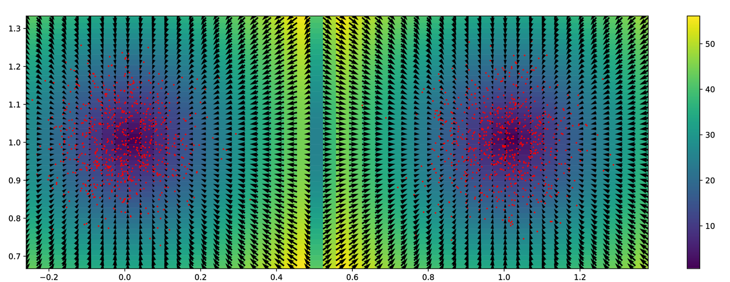
Theory

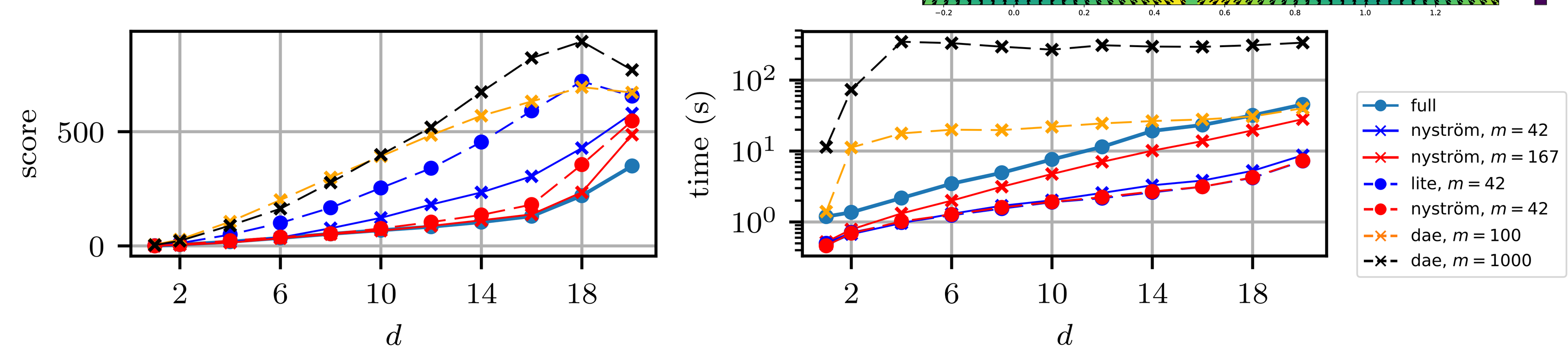
- Assume $p_0 = p_{f_0}$ for some $f_0 \in \mathcal{H}$; technical assumptions on $\mathcal{H}$, f_0
- θ a parameter depending on problem smoothness: worst case $\frac{1}{2}$, best $\frac{1}{3}$
- If we use "Nyström" with $m = \Omega(n^\theta \log n)$, $\lambda = n^{-\theta}$:
  - "Easy" problems: same convergence in J, $\mathcal{H}$, L_r, KL, Hellinger as [3]
  - "Hard" problems: same J convergence, others saturate slightly sooner
- Proof uses ideas from [2] for regression, but different decomposition:

$$f_\lambda^Y = \operatorname*{argmin}_{f \in \mathcal{H}_Y} J_\lambda(f); \quad \|f_{\lambda,n}^Y - f_0\|_{\mathcal{H}} \le \|f_{\lambda,n}^Y - f_\lambda^Y\|_{\mathcal{H}} + \|f_\lambda^Y - f_0\|_{\mathcal{H}}$$

Synthetic experiments

- Target: Gaussians centered on d vertices of d-dimensional hypercube
- Evaluate Fisher divergence $J(f)$:

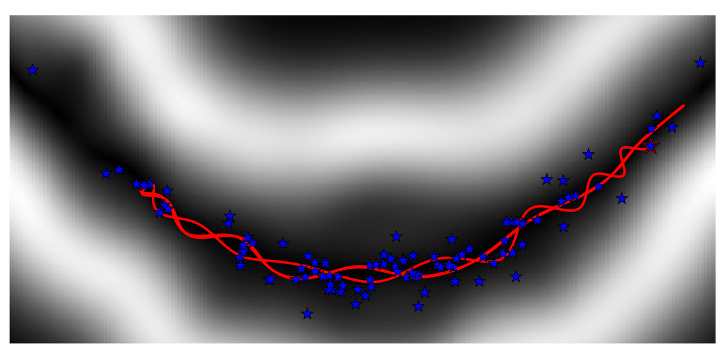




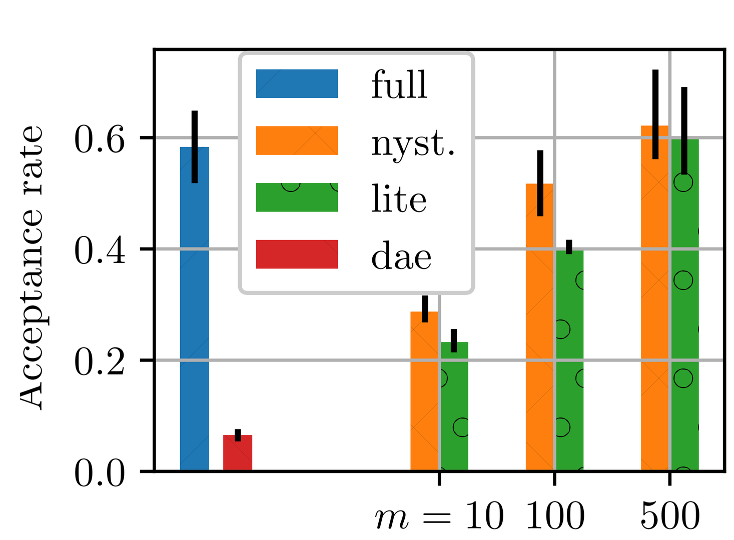
- Similar results for density around concentric rings

Approximate Hamiltonian Monte Carlo

- HMC uses $\nabla_x \log p(x)$, often more efficient
- Sometimes we can't get these gradients
  - e.g. marginalizing out hyperparameter choice for a GP classifier



- Kernel Adaptive HMC [4]:
  - Start with random walk MCMC
  - Estimate $\nabla_x \log p(x)$ from chain so far
  - Propose HMC trajectories with estimate
  - Metropolis rejection step accounts for errors in the proposed trajectories



Takeaways

- Flexible density modeling with kernel exponential families
- Nyström approximation: faster algorithm ($n^{\frac{5}{3}}$ to n^2) with same statistical guarantees as full-data fit (n^3)
- *Kernel Conditional Exponential Family*: less-smooth densities
- Open questions: kernel choice, theory for "lite" basis, misspecified case

References

[1] Canu and Smola. Kernel methods and the exponential family. *Neurocomputing* 2006.

[2] Rudi et al. Less is more: Nyström computational regularization. NIPS 2015.

[3] Sriperumbudur et al. Density estimation in infinite dimensional exponential families. JMLR 2017.

[4] Strathmann et al. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. NIPS 2015.