Efficient and principled score estimation

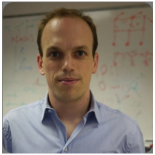
Nyström, kernels & exponential families

Heiko Strathmann, Gatsby unit

Gaussian process approximations workshop, Berlin

May 30, 2017

Joint work!

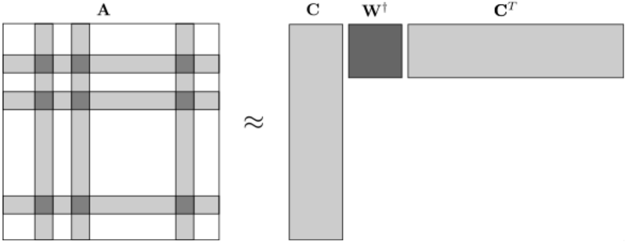
   

https://arxiv.org/abs/1705.08360

Frequentist guarantees for Bayesian methods?!?

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(X_i, \cdot) \qquad \tilde{f}(\cdot) = \sum_{\tilde{X}_i \in \text{subset}}^{m} \tilde{\alpha}_i k(\tilde{X}_i, \cdot)$$

- Nystrom approx. kernel matrix, Williams & Seeger (2000)



- Approx. basis, SoR/DTC, Quiñonero-Candela & Rasmussen (2005)

Frequentist guarantees for Bayesian methods?!?

At the core of approximate Gaussian processes:

1. Choose computationally effective basis
2. Construct stochastic process that correlates with GP
3. Optimize / minimize KL / variational compression

Titsias (2009), Hensman et al (2013), Matthews et al (2016), and many many more

Some questions:

- Consistency when $n, m \to \infty$?
- How many inducing points, $m\{\ll, <, =\}n$?
- Error guarantees for finite n, m?
- Trade-off computational savings / error?

Rudi & Rosasco (2015), Capponetto & DeVito (2007)

Context here: unnormalised density estimation

- Given iid samples from unknown p_0

$$X = \{X_b\}_{b\in[n]} \subset \mathbb{R}^d$$

- Want to fit model p such that (in some divergence)

$$p(x)/Z(p) \approx p_0(x)$$

- Related to learning the score itself,

$$\nabla_x \log p(x) \approx \nabla_x \log p_0(x)$$

- Not concerned with normaliser $Z(p)$
  - MCMC, gradient-free HMC, Strathmann et al (2015)
  - Deep / energy-based / autoencoders, LeCun et al (2006), Alain & Bengio (2014)
  - etc ...
- Here: $\log p \in \mathcal{H}$, reproducing kernel Hilbert space $\mathcal{H}$

RKHS exponential families

- Model family by Sriperumbudur et al (2014)

$$\left\{ p_f(x) := \exp\left(\underbrace{\langle f, k(x,\cdot)\rangle_{\mathcal{H}}}_{=f(x)} - \log Z(p_f) \right) \quad \middle| \quad Z(p_f) < \infty \right\}$$

- For certain k, dense in probability densities (KL, TV, ...)
- Crux: fitting – normalising constant is intractable

$$Z(p_f) = \int \exp(f(x))\mathrm{d}x$$

- Maximum likelihood ill-posed, Fukumizu (2009)

Score matching, Hyvärien (2005)

- Instead of ML, minimise Fisher score

$$J(f) = \frac{1}{2} \int p_0(x) \left\| \nabla_x \log p_f(x) - \nabla_x \log p_0(x) \right\|_2^2 \mathrm{d}x$$

- Remarkable: can rewrite and estimate from X

$$J(f) = \int p_0(x) \sum_{i=1}^{d} \left[\partial_i^2 \log p_f(x) + \frac{1}{2} \left(\partial_i \log p_f(x) \right)^2 \right] \mathrm{d}x + \mathrm{c}$$

- Representer theorem, closed form, (nd)-dim linear solve

$$\begin{aligned} f_{\lambda,n} &= \arg\min_{f \in \mathcal{H}} \hat{J}(f) + \frac{1}{2}\lambda \|f\|_{\mathcal{H}}^2 \\ &= \sum_{a=1}^{n} \sum_{i=1}^{d} \beta_{(a,i)} \partial_i k(X_a, \cdot) - \frac{1}{\lambda} \xi(\cdot) \\ \beta &= (G + \lambda I)^{-1} h \end{aligned}$$

Our algorithm: block-subsampled low rank

- Generalizes ealier approximations, Strathmann et al(2015)
- Nystrom basis $Y \subset X$ (can be any basis!)

$$
\begin{aligned}
f^m_{\lambda,n} &= \arg\min_{f \in \mathcal{H}_Y} \hat{J}(f) + \frac{1}{2}\lambda \|f\|^2_{\mathcal{H}} \\
&= \sum_{a=1}^{m} \sum_{i=1}^{d} (\beta_Y)_{(a,i)} \partial_i k(Y_a, \cdot) \\
\beta_Y &= -(\tfrac{1}{n} G_{XY}^{\mathsf{T}} G_{XY} + \lambda G_{YY})^{\dagger} h_Y
\end{aligned}
$$

Nystrom gains, here for block sub-sampled G matrix:

- Training time: $\mathcal{O}(nm^2)$ instead of $\mathcal{O}(n^3)$
- Training memory: $\mathcal{O}(nm)$ instead of $\mathcal{O}(n^2)$
- Evaluation of f: $\mathcal{O}(m)$ instead of $\mathcal{O}(n)$
- Only store basis Y after training

High level theory

Assumptions:

- Well specified case: assume $p_0 = p_{f_0}$ for some $f_0 \in \mathcal{H}$
- Given smoothnes parameter of p_0, $\theta \in [\frac{1}{3}, \frac{1}{2}]$
- ...
- Set number of Nystrom basis points $m = \Omega(n^\theta \log n)$

We can bound (as a function of n, m, etc.)

- Difference of estimated and true model in $\mathcal{H}$ (implies KL)

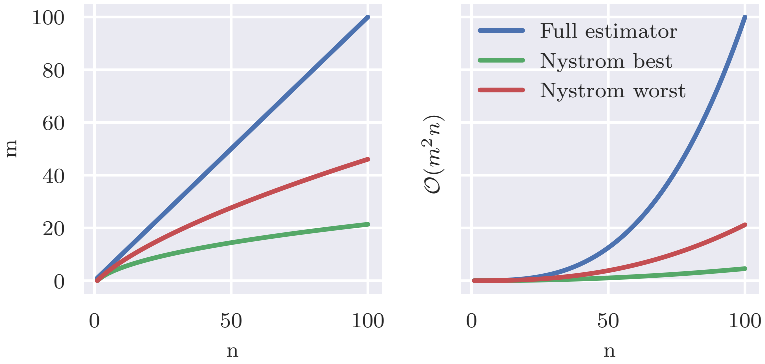
$$\|f^m_{\lambda,n} - f_0\|_{\mathcal{H}}$$

- Objective function (Fisher score) on test data

$$J(p_0 \| p_{f^m_{\lambda,n}})$$

The rates match those of the non-approximate estimator

Computational savings $m = \Omega(n^{\theta} \log n)$



The rates match those of the non-approximate estimator,
Sriperumbudur et al (2014), Rudi & Rosasco (2015),
Capponetto & DeVito (2007)

Proof 'outline'

- Decomposition

$$\|f^m_{\lambda,n} - f_0\|_{\mathcal{H}} \leq \underbrace{\|f^m_{\lambda,n} - f^m_\lambda\|_{\mathcal{H}}}_{\text{estimation error}} + \underbrace{\|f^m_\lambda - f_0\|_{\mathcal{H}}}_{\text{approx. error}}$$

Estimation error

- arises form finite samples
- independent of m
- decreases as $n \to \infty$
- increases as $\lambda \to 0$

Approximation error

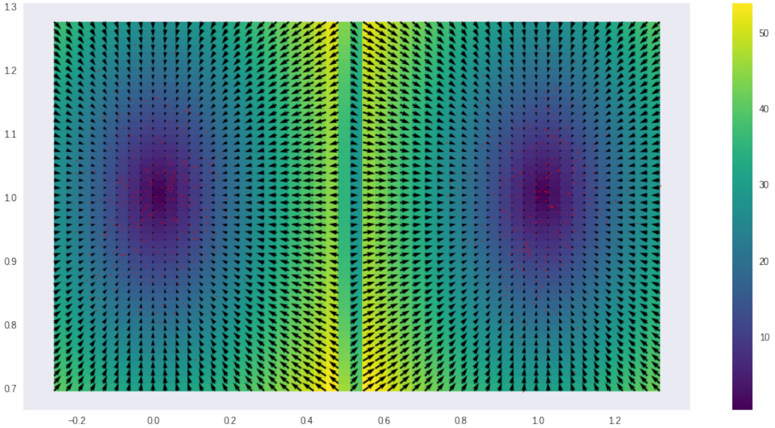
- arises from $Y \subset X$ and λ
- independent of n
- decreases as $m \to \infty$
- decreases as $\lambda \to 0$

Decay optimized for $\lambda = n^\theta$

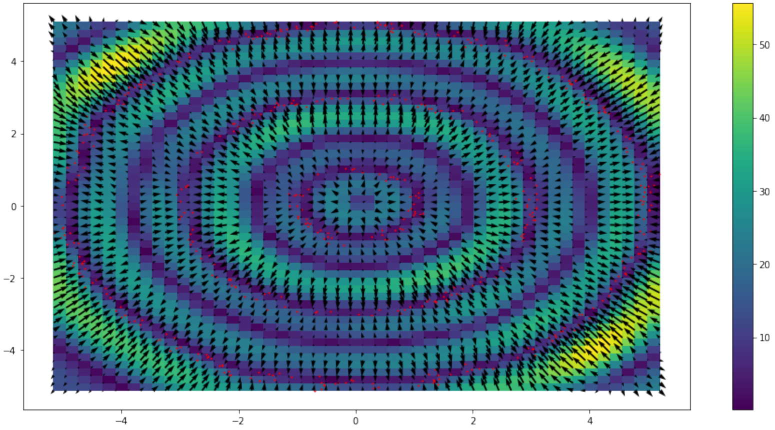
https://arxiv.org/abs/1705.08360

Convergence on synthetic data: Gaussian mixtures

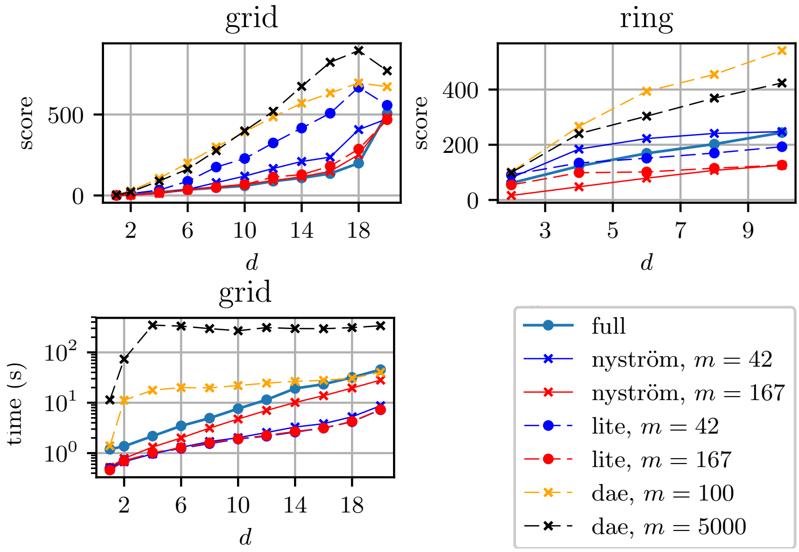
- d component Gaussian mixture in d dimensions
- centered on vertices of d-dimensional hypercube



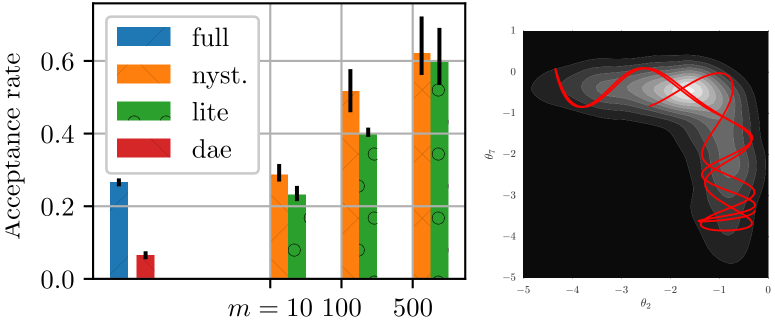
Convergence on synthetic data: ring



Convergence on synthetic data, n=500



Kernel HMC results for UCI glass dataset

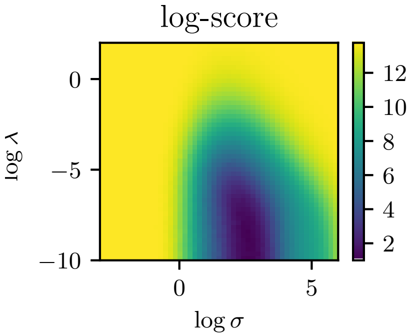


Hypothetical acceptance probabilies:

- ARD GP classification, gradient intractable
- fit models/score after (random walk) burn-in
- simulate HMC trajectory using fake gradient
- estimate HMC acceptance rate (peudo-marginal)

Parameter tuning: score on validation data

Regularization parameter λ and Gaussian kernel parameter σ



- Smooth surface, c.f. marginal likelihood
- Leave-one-out estimates possible in closed form
- Autodiff wrt. hyperparameters straight-forward
- 'Easily optimizable' (work in progress)

Summary

- Efficient
  - reduced computational costs
  - easy to implement
- Principled
  - explicit trade-offs between computation and n, m
  - error guarantees, matching the non-approximate version
- Practical
  - outperforms earlier kernel models
  - outperforms the autoencoder approach

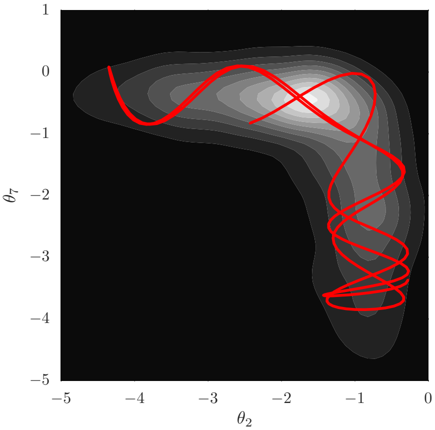
Points for discussion

GP community:

- Rudi & Rosasco (2015) and presented bounds apply to GP mean
- $\|f^m_{\lambda,n} - f_0\|_{\mathcal{H}}$ is related to the RMSE
- Frequentist guarantees for Bayesian methods – connect with the ELBO?

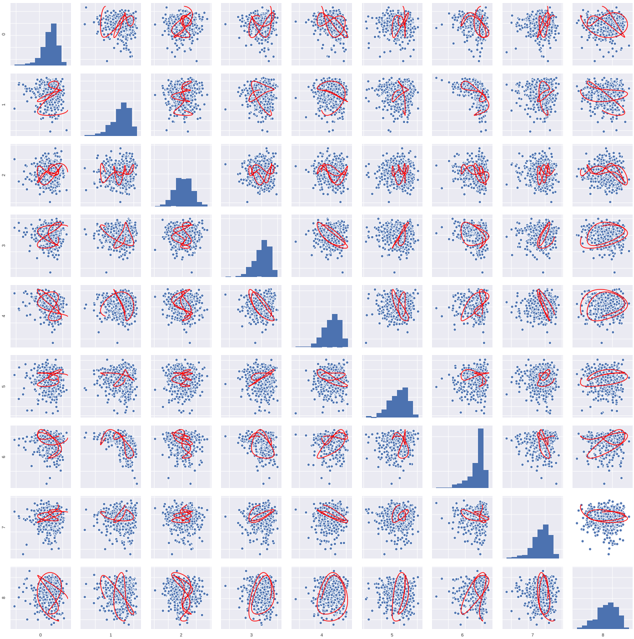
Thanks!

Motivation: gradient-free adaptive (kernel) HMC



- Posterior distribution over intractable GPs' parameters
- Strathmann et al (2015), Filippone & Girolami (2014)

Motivation: gradient-free adaptive (kernel) HMC

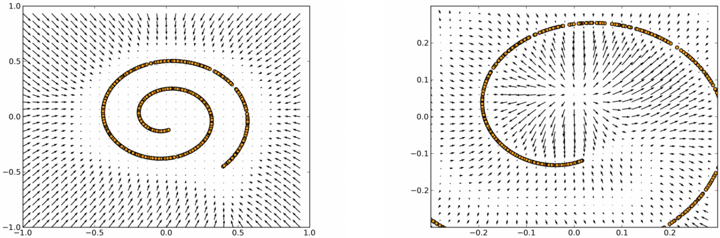


Denoising autoencoders, Alain & Bengio (2014)

- Autoencoder's reconstruction r_σ of noise corrupted input

$$(r_\sigma(x) - x)/\sigma^2 \approx \nabla_x \log p_0(x)$$

- Relies on universal approximation property
- No theory regarding number of hidden units, σ



(a) $r(x) - x$ vector field, acting as sink, zoomed out

(b) $r(x) - x$ vector field, close-up

Figure 5: The original 2-D data from the data generating density $p(x)$ is plotted along with the vector field defined by the values of $r(x) - x$ for trained auto-encoders (corresponding to the estimation of the score $\frac{\partial \log p(x)}{\partial x}$).