Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families

Heiko Strathmann*, Dino Sejdinovic+, Samuel Livingstone°, Zoltan Szabo*, Arthur Gretton*

*Gatsby Unit, University College London. +Department of Statistics, University of Oxford, °School of Mathematics, University of Bristol

Motivation: Hamiltonian Monte Carlo and Intractable Targets

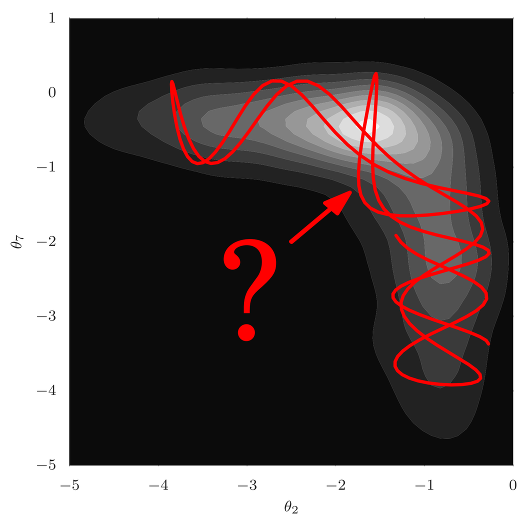
- Goal: Efficient sampling from density π on $\mathbb{R}^d$.
- HMC proposes distant moves with high acceptance probability.
- Given potential energy $U(q) = -\log \pi(q)$, sample auxiliary momentum $p \sim \exp(-K(p))$ and simulate for $t \in \mathbb{R}$ along Hamiltonian flow

$$\phi_t^H : (p, q) \mapsto (p^*, q^*)$$

of the joint log-density $H(p,q) = K(p) + U(q)$, using the operator

$$\frac{\partial K}{\partial p}\frac{\partial}{\partial q} - \frac{\partial U}{\partial q}\frac{\partial}{\partial p}$$

- Numerical simulation (i.e. leapfrog) depends on *gradient information*.
- Often *unavailable*, e.g. in Bayesian GP classification. More generally in Pseudo-Marginal MCMC [1] or Approximate Bayesian Computation [4].
- **Right**: Marginal hyper-parameters of a GP classifier. HMC dynamics?



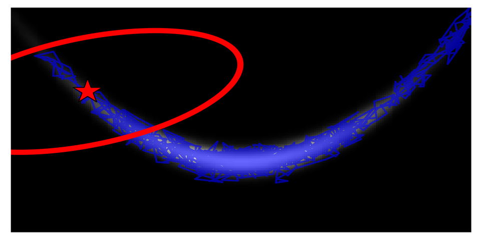
We want a HMC sampler that automatically learns gradients.

So far: (Kernel) Adaptive Metropolis-Hastings

Idea: use history of Markov chain to learn target structure.

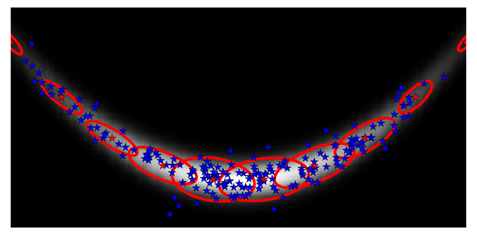
Adaptive Metropolis-Hastings [2]

- Learns *global* linear covariance.
- Pro: Automatically learns proposal scaling, fast.
- Con: Local steps, does not work well on non-linear targets.



Kernel Adaptive Metropolis Hastings [5]

- Learns covariance in RKHS.
- Pro: *Locally* aligns to (non-linear) target covariance, gradient free.
- Con: Local steps, random walk.



Can we combine 'global' and 'non-linear' – without gradients?

Hamiltonian Monte Carlo with kernel induced potential energy

- Learn gradient 'surrogate' model $\nabla U_k \approx \nabla U = -\nabla \log \pi$ from Markov chain history $\{x_i\}_{i=1}^t$.
- Replace $\frac{\partial U}{\partial q}$ by $\frac{\partial U_k}{\partial q}$; gives kernel induced Hamiltonian flow $\phi_t^{H_k} : (p,q) \mapsto (p_k^*, q_k^*)$
- $\phi_t^{H_k}$ can be simulated using the operator

$$\frac{\partial K}{\partial p}\frac{\partial}{\partial q} - \frac{\partial U_k}{\partial q}\frac{\partial}{\partial p}$$

- Accept using *true* Hamiltonian (depends on U but *not* on ∇U) with probability

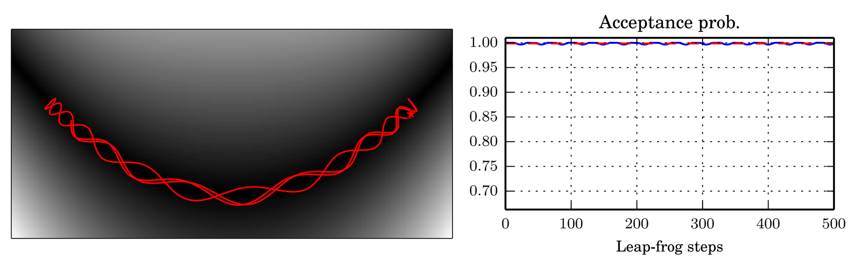
$$\min\left[1, \exp\left(-H\left(p_k^*, q_k^*\right) + H(p,q)\right)\right]$$

- Corrects for both leap-frog error *and* surrogate induced Hamiltonian flow error $\Rightarrow$ Asymptotically correct.
- **Note**: $\exp(U(q))$ can be replaced with unbiased estimator, c.f. Pseudo-Marginal MCMC.

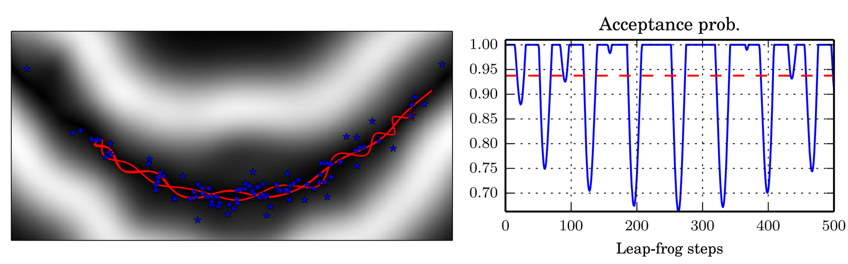
Key quantity: average gradient error $\int \pi(x) \|\nabla U(x) - \nabla U_k(x)\|_2^2 dx$

Illustration of kernel induced Hamiltonian flow

- Standard HMC dynamics using ∇U (plot shows gradient norm $\|\nabla U\|$).



- Dynamics on kernel surrogate ∇U_k, fitted from samples.



We need an expressive yet tractable model.

Infinite dimensional exponential families [6]

(Unnormalised) exponential family model in a RKHS:

$$\text{const} \times \pi(x) \approx \exp\left(\langle f, k(x,\cdot)\rangle_{\mathcal{H}} - A(f)\right)$$

- Sufficient statistics: feature map $k(\cdot, x) \in \mathcal{H}$, satisfies $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$.
- Natural parameters: $f \in \mathcal{H}$.
- Normalising constant $A(f)$ is intractable.

The model is

- dense in continuous densities on compact domains (in TV, KL, etc.),
- relatively robust to increasing dimensions, as opposed to e.g. KDE.

How to learn f from samples without access to $A(f)$?

Score matching [3]

- Allows estimation of unnormalised density models from samples.
- Minimises *Fisher divergence* (precisely the average gradient error):

$$J(f) = \frac{1}{2}\int \pi(x) \|\nabla f(x) - \nabla \log \pi(x)\|_2^2 dx$$

- Possible *without* accessing $\nabla \log \pi(x)$, and accessing $\pi(x)$ only through samples: $\mathbf{x} := \{x_i\}_{i=1}^t$

$$\hat{J}(f) = \frac{1}{|\mathbf{x}|}\sum_{x \in \mathbf{x}} \sum_{\ell=1}^d \left[\frac{\partial^2 f(x)}{\partial x_\ell^2} + \frac{1}{2}\left(\frac{\partial f(x)}{\partial x_\ell}\right)^2\right]$$

Expensive: Closed form full solution requires solving $(td+1)$-dimensional linear system.

Approximation I: KMC Lite

Assume that the model takes the form (Gaussian kernel k with bandwidth σ)

$$f_{\mathtt{lite}}(x) = \sum_{i=1}^n \alpha_i k(z_i, x)$$

- $\mathbf{z} \subseteq \mathbf{x}$ is a random sub-sample, $\alpha \in \mathbb{R}^n$ are real valued parameters.
- Solution $f_{\mathtt{lite}}$ lies in smaller RKHS sub-space than original model, yet grows with $n \ll t$.
- Compute α from linear system

$$\hat{\alpha}_\lambda = -\frac{\sigma}{2}(C + \lambda I)^{-1} b$$

where $C \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n$ depend on kernel matrix, and $\lambda > 0$.
- Costs $\mathcal{O}(n^3 + n^2 d)$. Can further reduce with low-rank approximations and conjugate gradient.

Approximation II: KMC finite

Assume that the model takes the form

$$f_{\mathtt{finite}}(x) = \theta^\top \phi_x$$

- $\phi_x \in \mathbb{R}^m$ is approximate feature map such that $\phi_x^\top \phi_y \approx k(x,y)$, c.f. *Random Fourier Features*.
- $\theta \in \mathbb{R}^m$ can be computed from

$$\hat{\theta}_\lambda := (C + \lambda I)^{-1} b$$

where

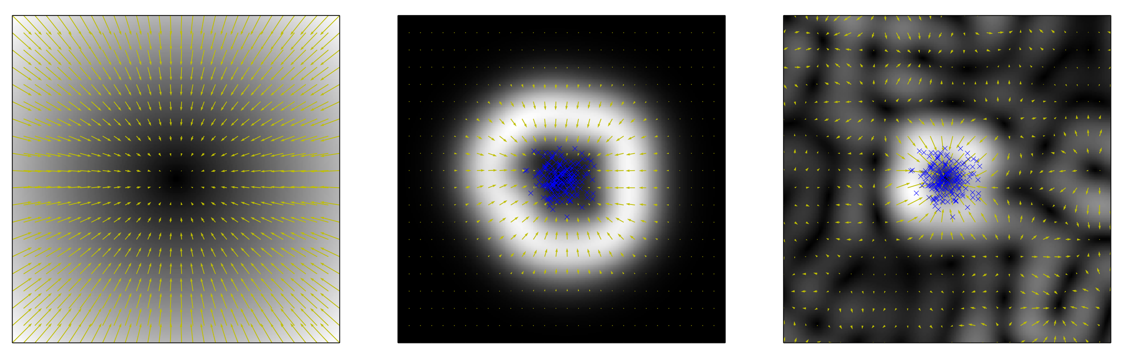
$$b := -\frac{1}{n}\sum_{i=1}^t \sum_{\ell=1}^d \ddot{\phi}_{x_i}^\ell \in \mathbb{R}^m \qquad C := \frac{1}{n}\sum_{i=1}^t \sum_{\ell=1}^d \dot{\phi}_{x_i}^\ell \left(\dot{\phi}_{x_i}^\ell\right)^T \in \mathbb{R}^{m \times m}$$

where $\dot{\phi}_x^\ell := \frac{\partial}{\partial x_\ell}\phi_x$ and $\ddot{\phi}_x^\ell := \frac{\partial^2}{\partial x_\ell^2}\phi_x$ and $\lambda > 0$.

- C, b are running averages. *On-line updates* cost $\mathcal{O}(dm^2)$.

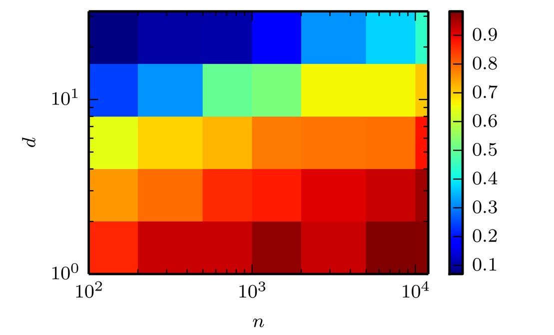
Lite vs. Finite: geometric ergodicity & the tails

- KMC lite is geometrically ergodic on log-concave targets (fast convergence).
- KMC finite updates fast and uses *all* Markov chain history. Caveat: need to initialise correctly.
- Gradient norm of



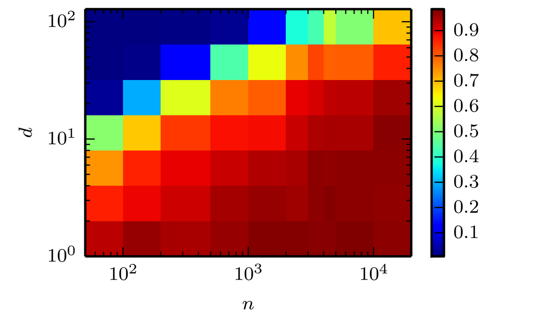
Stability in growing dimensions

- Fit surrogate on n oracle samples, increase d and n.
- Compute acceptance rate along random HMC trajectories.
- Small step-size, optimal value is 1.
- **Red**: KMC efficient, **blue**: KMC inefficient.



A challenging Gaussian target (**top**):

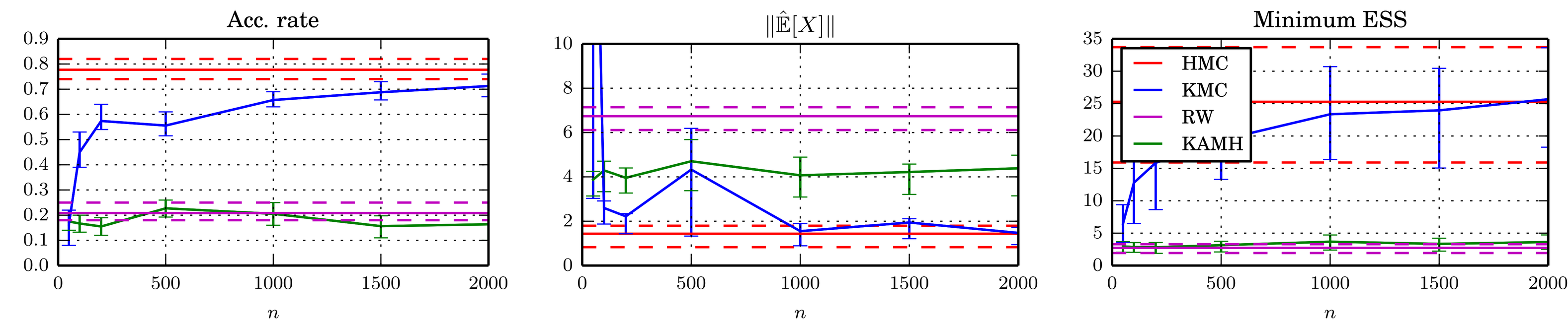
- Eigenvalues: $\lambda_i \sim \text{Exp}(1)$.
- Covariance: $\text{diag}(\lambda_1, \ldots, \lambda_d)$, randomly rotate.
- Use Rational Quadratic kernel to account for resulting highly 'non-singular' length-scales.
- KMC scales up to $d \approx 30$.



An easy, isotropic Gaussian target (**bottom**):

- More smoothness allows KMC to scale up to $d \approx 100$.

Mixing on synthetic 8-dimensional Banana [5]



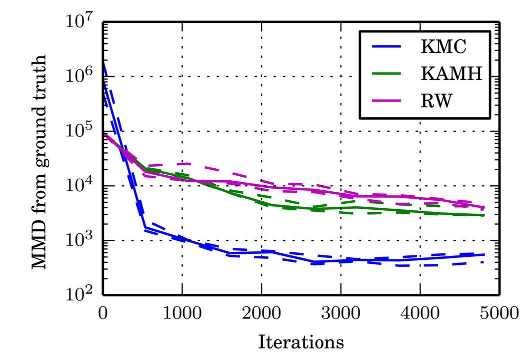
KMC behaves like HMC as number n of oracle samples increases.

Gaussian Process Classification on UCI data

- Standard GPC model

$$p(\mathbf{f}, \mathbf{y}, \theta) = p(\theta)p(\mathbf{f}|\theta)p(\mathbf{y}|\mathbf{f})$$

where $p(\mathbf{f}|\theta)$ is a GP and with a sigmoidal likelihood $p(\mathbf{y}|\mathbf{f})$.
- Goal: sample from $p(\theta|\mathbf{y}) \propto p(\theta)p(\mathbf{y}|\theta)$.
- Unbiased estimate of $\hat{p}(\mathbf{y}|\theta)$ via importance sampling.
- No access to likelihood or gradient.



Significant mixing improvements over state-of-the-art.

Approximate Bayesian Computation on a Skew-Normal model

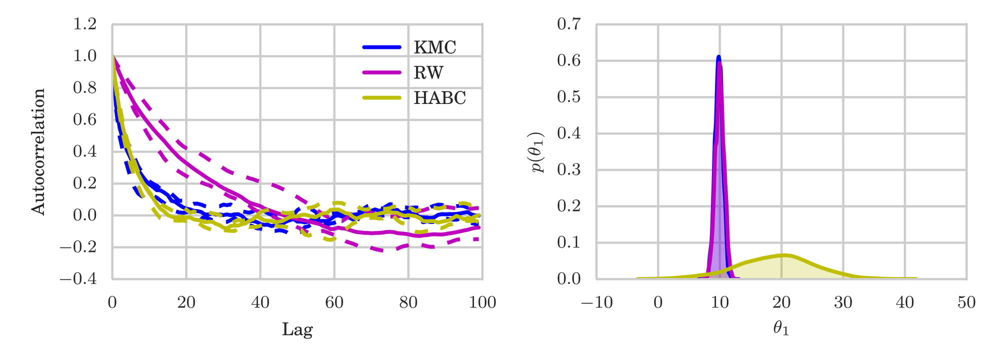
- Likelihood free MCMC for ABC via simulation from likelihood.
- Can fit (Gaussian) synthetic likelihoods.
- This often induces bias, simple example:

$$p(\mathbf{y}|\theta) = 2\mathcal{N}(\mathbf{y}|\theta, I)\,\Phi\left(\alpha^\top \mathbf{y}\right)$$

with Gaussian CDF Φ and skewness $\alpha = 10 \cdot \mathbf{1}^\top$.

Compared to Hamiltonian ABC (gradients by stochastic finite differences):

- KMC uses surrogate for ABC *posterior*.
- No synthetic likelihood.
- No stochastic gradients.



No additional bias and reduced number of likelihood simulations.

References

[1] C. Andrieu and G.O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, April 2009.

[2] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, December 2008.

[3] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *JMLR*, 6:695–709, 2005.

[4] E. Meeds, R. Leenders, and M. Welling. Hamiltonian ABC. In *UAI*, 2015.

[5] D. Sejdinovic, H. Strathmann, M. Lomeli, C. Andrieu, and A. Gretton. Kernel Adaptive Metropolis-Hastings. In *ICML*, 2014.

[6] B. Sriperumbudur, K. Fukumizu, R. Kumar, A. Gretton, and A. Hyvärinen. Density Estimation in Infinite Dimensional Exponential Families. *arXiv preprint arXiv:1312.3516*, 2014.

This work was funded by the Gatsby Charitable Foundation.

Code: `https://github.com/karlnapf/kernel_hmc`

Contact: `heiko.strathmann@gmail.com`