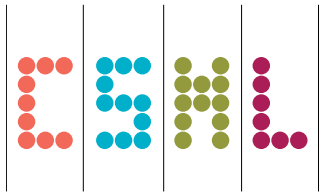
Adaptive HMC via the Infinite Exponential Family

Arthur Gretton

*Gatsby Unit, CSML, University College London

RegML, 2017



Using samples to compute expectations

- We have a density of the form

$$p(x) = \frac{\pi(x)}{Z} \qquad Z = \int \pi(x) dx$$

  Z often impractical to compute
- **Goal:** to compute expectations of functions,

$$\mathbb{E}_p[f(x)] = \int f(x) p(x) dx$$

Using samples to compute expectations

- We have a density of the form

$$p(x) = \frac{\pi(x)}{Z} \qquad Z = \int \pi(x) dx$$

  Z often impractical to compute

- **Goal:** to compute expectations of functions,

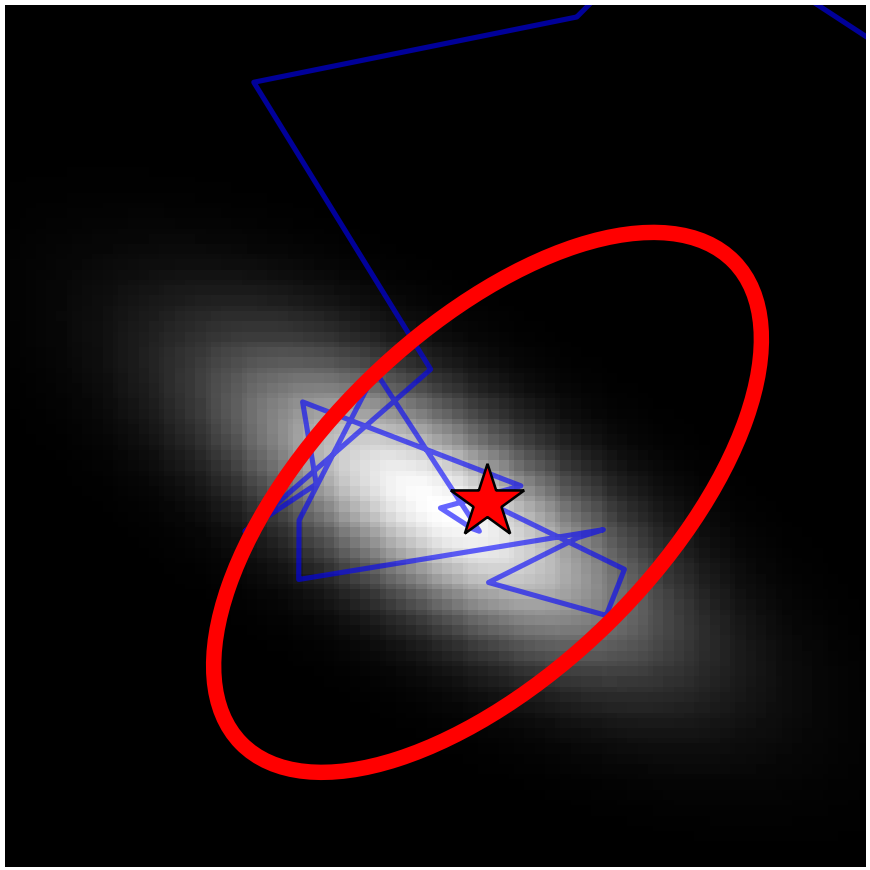
$$\mathbb{E}_p[f(x)] = \int f(x) p(x) dx$$

- Given **samples** $\{x_i\}_{i=1}^n$ with distribution $p(x)$,

$$\widehat{\mathbb{E}}_p[f(x)] = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

Metropolis-Hastings MCMC

A visual guide...



Metropolis-Hastings MCMC

- Unnormalized target $\pi(x) \propto p(x)$
- Generate Markov chain with invariant distribution p
  - Initialize $x_0 \sim p_0$
  - At iteration $t \geq 0$, propose to move to state $x' \sim q(\cdot|x_t)$
  - Accept/Reject proposals based on ratio

$$x_{t+1} = \begin{cases} x', & \text{w.p. } \min\left\{1, \frac{\pi(x')q(x_t|x')}{\pi(x_t)q(x'|x_t)}\right\}, \\ x_t, & \text{otherwise.} \end{cases}$$

Metropolis-Hastings MCMC

- Unnormalized target $\pi(x) \propto p(x)$
- Generate Markov chain with invariant distribution p
  - Initialize $x_0 \sim p_0$
  - At iteration $t \geq 0$, propose to move to state $x' \sim q(\cdot|x_t)$
  - Accept/Reject proposals based on ratio

$$x_{t+1} = \begin{cases} x', & \text{w.p. } \min\left\{1, \frac{\pi(x')q(x_t|x')}{\pi(x_t)q(x'|x_t)}\right\}, \\ x_t, & \text{otherwise.} \end{cases}$$

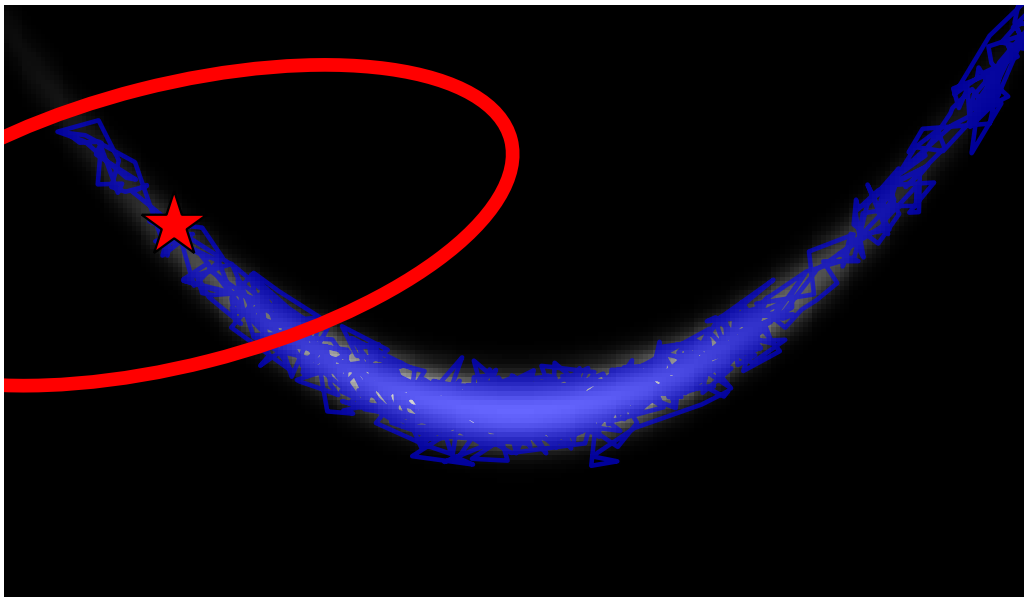
- What proposal $q(\cdot|x_t)$?
  - Too narrow or broad: $\rightarrow$ slow convergence
  - Does not conform to support of target $\rightarrow$ slow convergence

Adaptive MCMC

- **Adaptive Metropolis** (Haario, Saksman & Tamminen, 2001): Update proposal $q_t(\cdot|x_t) = \mathcal{N}(x_t, \nu^2\hat{\Sigma}_t)$, using estimates of the target covariance

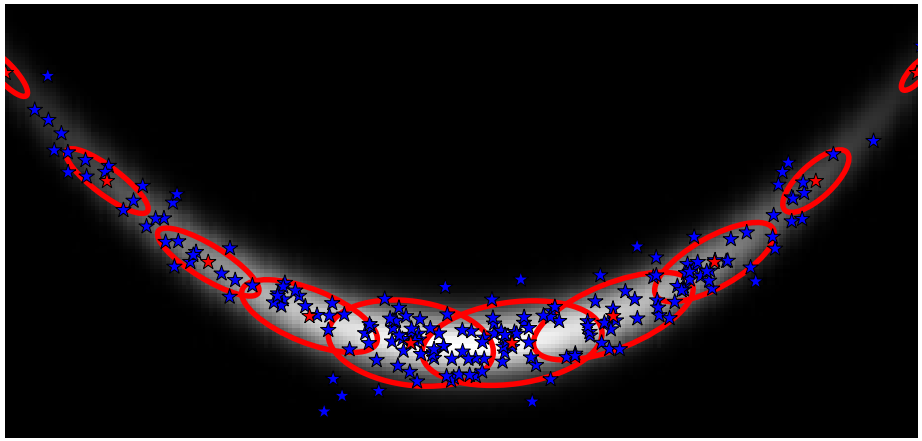
Adaptive MCMC

- **Adaptive Metropolis** (Haario, Saksman & Tamminen, 2001): Update proposal $q_t(\cdot|x_t) = \mathcal{N}(x_t, \nu^2\hat{\Sigma}_t)$, using estimates of the target covariance



Locally miscalibrated for *strongly non-linear targets*: directions of large variance depend on the current location

Alternative adaptive sampler: the Kameleon



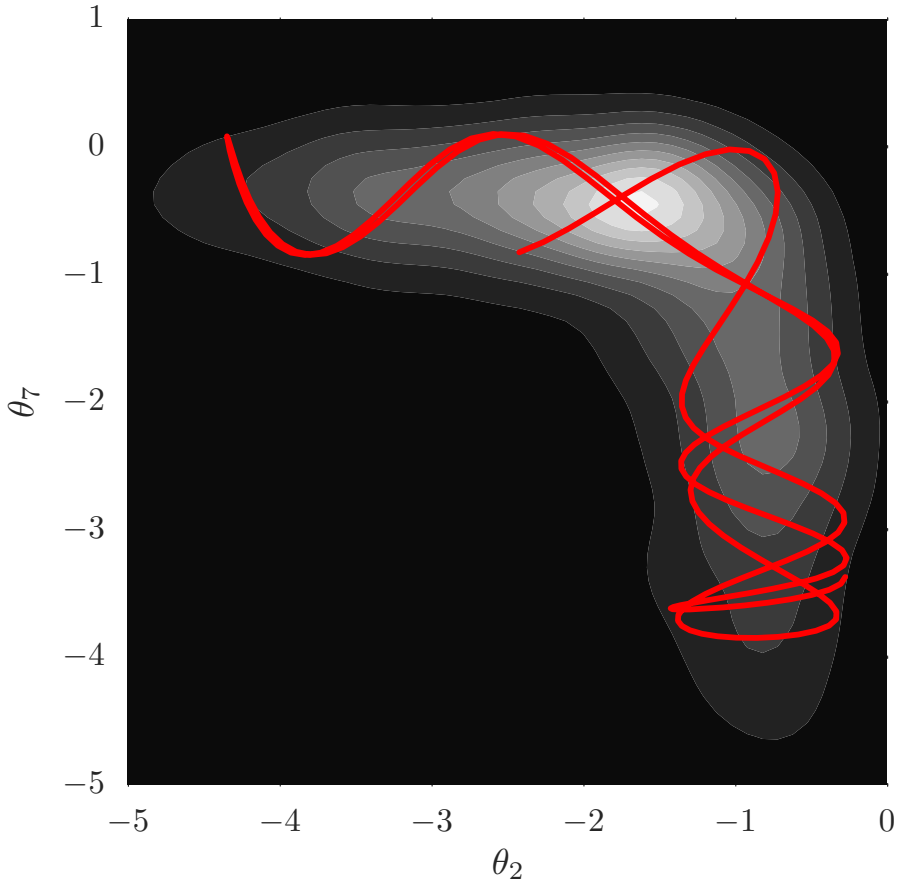
- Idea: fit Gaussian in feature space, take local steps in directions of max. principal components.

D. Sejdinovic, H. Strathmann, M. Lomeli, C. Andrieu, and A. Gretton, ICML 2014

Hamiltonian Monte Carlo

- HMC: distant moves, high acceptance probability.
- Potential energy $U(x) = -\log \pi(x)$, auxiliary momentum $p \sim \exp(-K(p))$, simulate for $t \in \mathbb{R}$ along Hamiltonian flow of $H(p, x) = K(p) + U(x)$, using operator

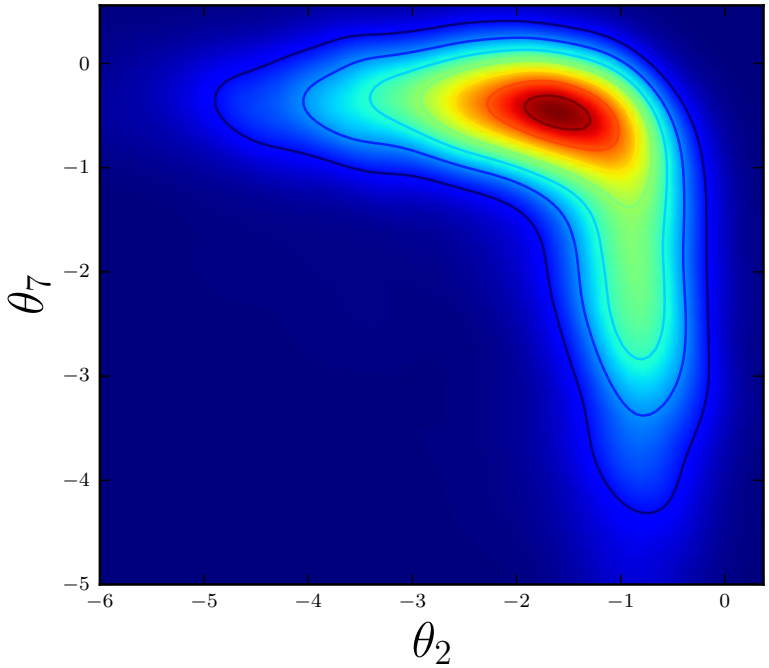
$$\frac{\partial K}{\partial p}\frac{\partial}{\partial x} - \frac{\partial U}{\partial x}\frac{\partial}{\partial p}$$



- Numerical simulation (i.e. leapfrog) depends on *gradient information*.

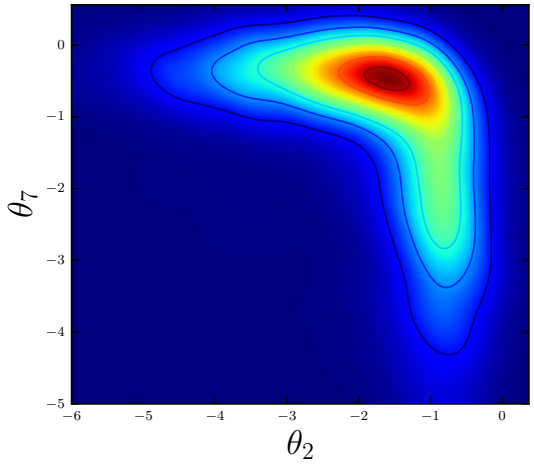
Intractable & Non-linear Target in GPC

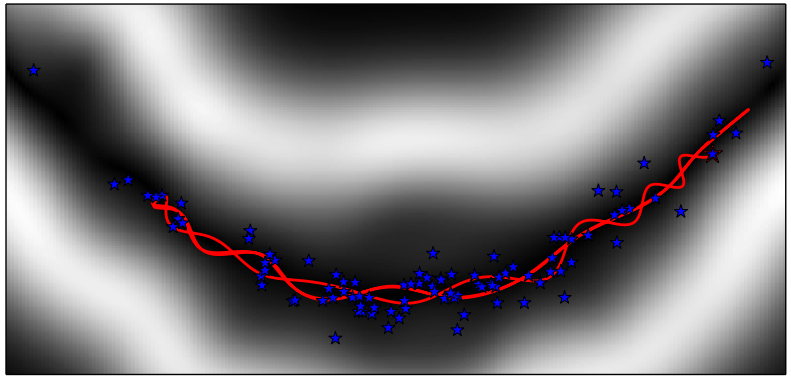
- Sliced posterior over hyperparameters of a Gaussian Process classifier on UCI Glass dataset obtained using Pseudo-Marginal MCMC



Can you learn an HMC sampler?

Outline for remainder of talk





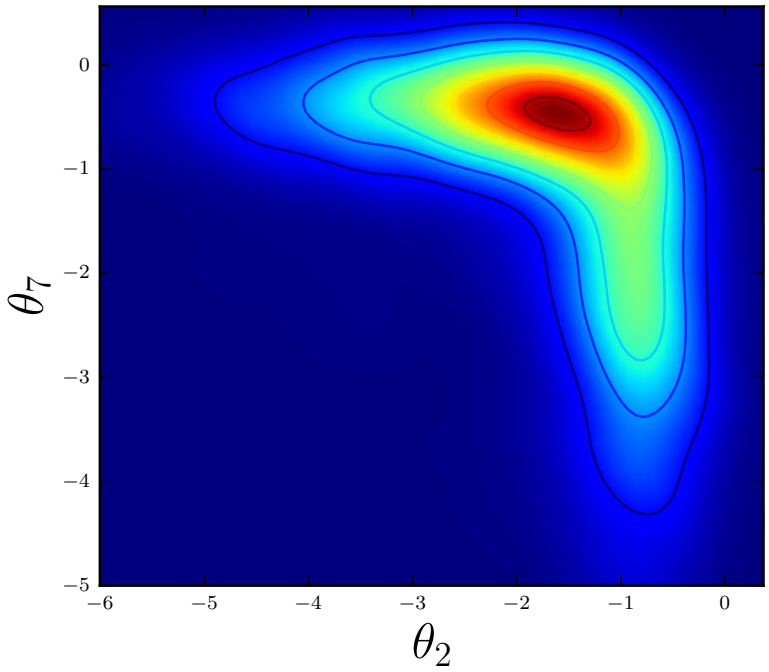
Infinite dimensional exponential family (Sriperumbudur et al. 2014)

- Exponential family with RKHS-valued natural parameter
- Learned via *score matching*, no log-partition function

Kernel Adaptive Hamiltonian Monte Carlo (Strathmann et al. 2015)

- *Global* estimate of gradient of log target density from prev. samples
- Mixing performance close to ideal "known density" HMC

Infinite dimensional exponential family density estimator



Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Revant Kumar, and Aapo Hyvarinen, JMLR 2017, to appear

(slides adapted from Bharath's talk)

The Exponential Family of Distributions

- Natural form:

$$p_\theta(x) = q_0(x) e^{\theta^T T(x) - A(\theta)}$$

where

- $\theta \in \Theta \subset \mathbb{R}^m$ (natural parameter)
- q_0: probability density defined over $\Omega \subset \mathbb{R}^d$
- $A(\theta)$: log-partition function

$$A(\theta) = \log \int e^{\theta^T T(x)} q_0(x)\, dx$$

- $T(x)$: sufficient statistic

- Includes many commonly used distributions
  - Normal, Binomial, Poisson, Exponential, ...

Infinite Dimensional Generalization

$$\mathcal{P} = \left\{ p_f(x) = e^{f(x) - A(f)} q_0(x),\ x \in \Omega\ :\ f \in \mathcal{F} \right\}$$

where

$$\mathcal{F} = \left\{ f \in \mathcal{H} : A(f) = \log \int e^{f(x)} q_0(x)\, dx < \infty \right\}$$

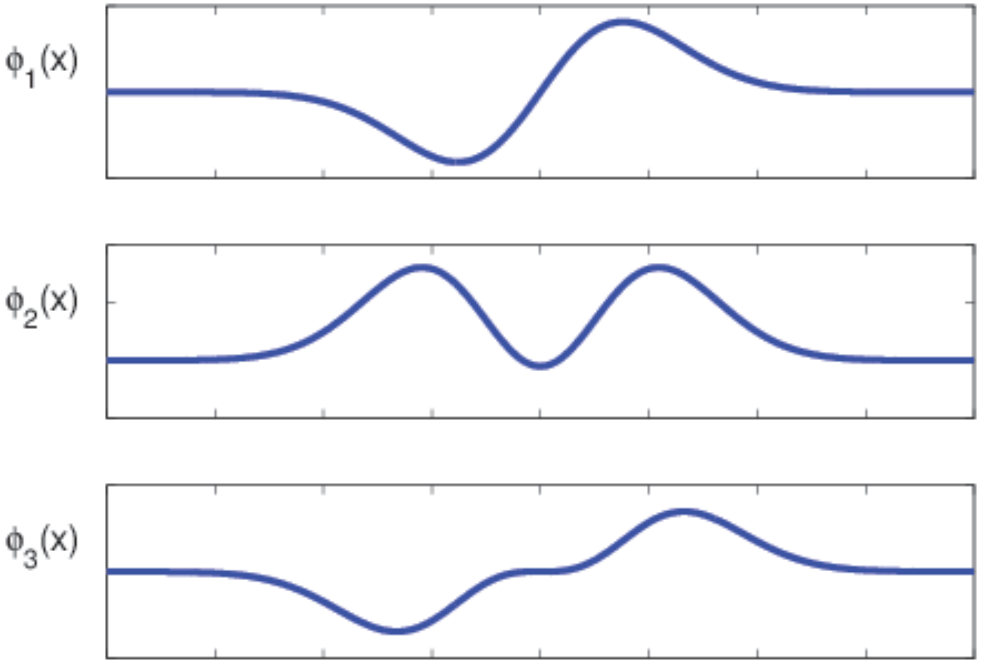
- (Canu and Smola, 2005; Fukumizu, 2009): $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

Reproducing kernel Hilbert space

Exponentiated quadratic kernel,

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) = \sum_{i=1}^{\infty} \phi_i(x)\phi_i(x')$$

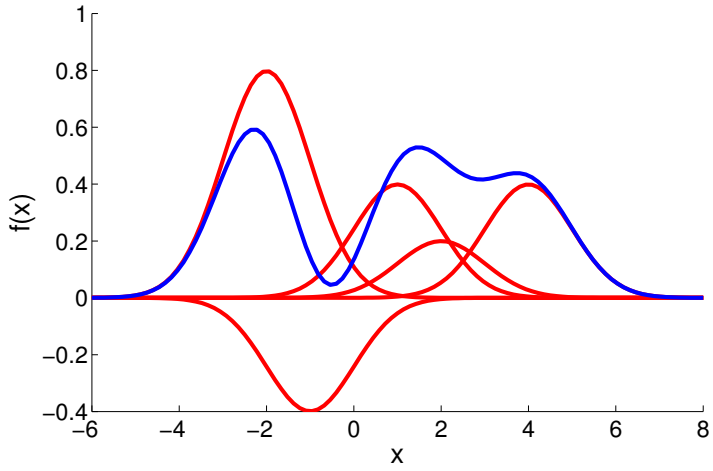
$$f(x) = \sum_{i=1}^{\infty} f_i \phi_i(x) \qquad \sum_{i=1}^{\infty} f_i^2 < \infty.$$



Reproducing kernel Hilbert space

Function with exponentiated quadratic kernel:

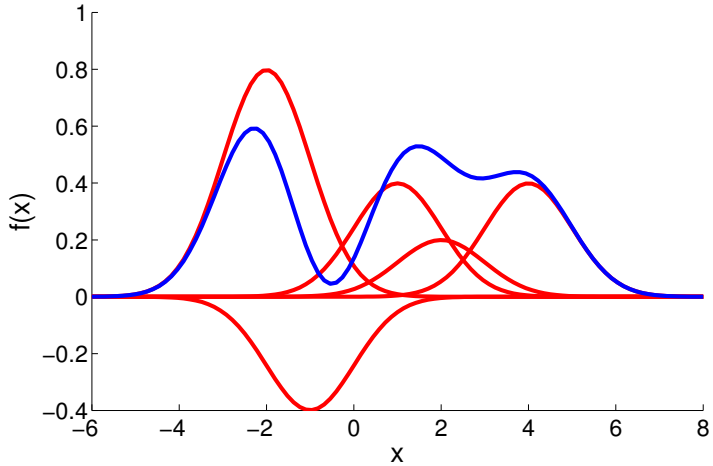
$$
\begin{aligned}
f(x) &:= \sum_{i=1}^{m} \alpha_i k(x_i, x) \\
&= \sum_{i=1}^{m} \alpha_i \langle \phi(x_i), \phi(x) \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}}
\end{aligned}
$$



Reproducing kernel Hilbert space

Function with exponentiated quadratic kernel:

$$
\begin{aligned}
f(x) &:= \sum_{i=1}^{m} \alpha_i k(x_i, x) \\
&= \sum_{i=1}^{m} \alpha_i \langle \phi(x_i), \phi(x) \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}} \\
&= \sum_{\ell=1}^{\infty} f_\ell \phi_\ell(x) \\
&= \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}}
\end{aligned}
$$



$f_\ell := \sum_{i=1}^{m} \alpha_i \phi_\ell(x_i)$

Possible to write functions of infinitely many features!

RKHS-Based Exponential Family

- $\mathcal{H}$ is an RKHS:

$$\mathcal{P} = \left\{ p_f(x) = e^{\langle f, \phi(x) \rangle_{\mathcal{H}} - A(f)} q_0(x), \ x \in \Omega, f \in \mathcal{F} \right\}$$

where

$$\mathcal{F} = \left\{ f \in \mathcal{H} : A(f) = \log \int e^{f(x)} q_0(x)\, dx < \infty \right\}.$$

- Finite dimensional RKHS: one-to-one correspondence between finite dimensional exponential family and RKHS.
- $T(x) \rightsquigarrow k(x, y) = \langle T(x), T(y) \rangle$. Similarly, $k(x, y) = \langle \Phi(x), \Phi(y) \rangle \rightsquigarrow \Phi(x)$.

Examples

Exponential: $\Omega = \mathbb{R}_{++}$, $k(x, y) = xy$.

Normal: $\Omega = \mathbb{R}$, $k(x, y) = xy + x^2 y^2$.

Beta: $\Omega = (0, 1)$, $k(x, y) = \log x \log y + \log(1 - x) \log(1 - y)$.

Gamma: $\Omega = \mathbb{R}_{++}$, $k(x, y) = \log x \log y + xy$.

Inverse Gaussian: $\Omega = \mathbb{R}_{++}$, $k(x, y) = xy + \frac{1}{xy}$.

Poisson: $\Omega = \mathbb{N} \cup \{0\}$, $k(x, y) = xy$, $q_0(x) = (x!\, e)^{-1}$.

Geometric: $\Omega = \mathbb{N} \cup \{0\}$, $k(x, y) = xy$, $q_0(x) = 1$.

Binomial: $\Omega = \{0, \ldots, m\}$, $k(x, y) = xy$, $q_0(x) = 2^{-m} \binom{m}{c}$.

Problem: Given random samples, $X_1, \ldots, X_n$ drawn i.i.d. from an unknown density, $p_0 := p_{f_0} \in \mathcal{P}$, estimate p_0.

Maximum Likelihood Estimation

$$
\begin{aligned}
f_{ML} &= \arg\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log p_f(X_i) \\
&= \arg\max_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i) - n \log \int e^{f(x)} q_0(x)\, dx.
\end{aligned}
$$

Maximum Likelihood Estimation

$$
\begin{aligned}
f_{ML} &= \arg\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log p_f(X_i) \\
&= \arg\max_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i) - n \log \int e^{f(x)} q_0(x)\, dx.
\end{aligned}
$$

Solving the above yields that f_{ML} satisfies

$$
\frac{1}{n} \sum_{i=1}^{n} \phi(x_i) = \int \phi(x) p_{f_{ML}}(x)\, dx
$$

Can we solve this?

Solving max. likelihood equations

- Finite dimensional case: Normal distribution $\mathcal{N}(\mu, \sigma)$

$$\phi(x) = \begin{bmatrix} x & x^2 \end{bmatrix}^\top$$

- Max. likelihood equations give

$$\frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} x_i & x_i^2 \end{bmatrix}^\top = \int \begin{bmatrix} x & x^2 \end{bmatrix}^\top p_{f_{ML}}(x)\, dx = \begin{bmatrix} \mu_{ML} & (\sigma_{ML}^2 + \mu_{ML}^2) \end{bmatrix}^\top$$

- System of likelihood equations: solvable

Solving max. likelihood equations

- Finite dimensional case: Normal distribution $\mathcal{N}(\mu, \sigma)$

$$\phi(x) = \begin{bmatrix} x & x^2 \end{bmatrix}^\top$$

- Max. likelihood equations give

$$\frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} x_i & x_i^2 \end{bmatrix}^\top = \int \begin{bmatrix} x & x^2 \end{bmatrix}^\top p_{f_{ML}}(x)\, dx = \begin{bmatrix} \mu_{ML} & (\sigma_{ML}^2 + \mu_{ML}^2) \end{bmatrix}^\top$$

- System of likelihood equations: solvable

- Infinite dimensional case, characteristic kernel: ill-posed!
- (Fukumizu, 2009): a sieves method involving pseudo-MLE by restricting $\mathcal{P}$ to a series of finite dimensional submanifolds, which enlarge as the sample size increases.

Score matching (general version)

- Score matching (Hyvärinen, 2005) since MLE can be intractable even in finite dimensions if $A(\theta)$ is not easily computable.
- Assuming p_f to be differentiable (w.r.t. x) and $\int p_0(x)\|\nabla_x \log p_f(x)\|^2 \, dx < \infty$, $\forall\, \theta \in \Theta$:

$$D_F(p_0 \| p_f) := \frac{1}{2} \int p_0(x) \, \|\nabla_x \log p_0(x) - \nabla_x \log p_f(x)\|^2 \, dx$$

Score matching: 1-D proof

$$D_F(p_0, p_f)$$

$$= \frac{1}{2} \int_a^b p_0(x) \left(\frac{d \log p_0(x)}{dx} - \frac{d \log p_f(x)}{dx} \right)^2 dx$$

Score matching: 1-D proof

$$
\begin{aligned}
& D_F(p_0, p_f) \\
&= \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_0(x)}{dx} - \frac{d\log p_f(x)}{dx}\right)^2 dx \\
&= \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_0(x)}{dx}\right)^2 dx + \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_f(x)}{dx}\right)^2 dx \\
&\quad - \int_a^b p_0(x)\left(\frac{d\log p_f(x)}{dx}\right)\left(\frac{d\log p_0(x)}{dx}\right) dx
\end{aligned}
$$

Score matching: 1-D proof

$$D_F(p_0, p_f)$$

$$= \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_0(x)}{dx} - \frac{d\log p_f(x)}{dx}\right)^2 dx$$

$$= \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_0(x)}{dx}\right)^2 dx + \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_f(x)}{dx}\right)^2 dx$$

$$- \int_a^b p_0(x)\left(\frac{d\log p_f(x)}{dx}\right)\left(\frac{d\log p_0(x)}{dx}\right) dx$$

Final term:

$$\int_a^b p_0(x)\left(\frac{d\log p_f(x)}{dx}\right)\left(\frac{d\log p_0(x)}{dx}\right) dx$$

Score matching: 1-D proof

$$D_F(p_0, p_f)$$

$$= \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_0(x)}{dx} - \frac{d\log p_f(x)}{dx}\right)^2 dx$$

$$= \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_0(x)}{dx}\right)^2 dx + \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_f(x)}{dx}\right)^2 dx$$

$$- \int_a^b p_0(x)\left(\frac{d\log p_f(x)}{dx}\right)\left(\frac{d\log p_0(x)}{dx}\right) dx$$

Final term:

$$\int_a^b p_0(x)\left(\frac{d\log p_f(x)}{dx}\right)\left(\frac{d\log p_0(x)}{dx}\right) dx$$

$$= \int_a^b \cancel{p_0(x)}\left(\frac{d\log p_f(x)}{dx}\right)\left(\frac{1}{\cancel{p_0(x)}}\frac{dp_0(x)}{dx}\right) dx$$

Score matching: 1-D proof

$$D_F(p_0, p_f)$$

$$= \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_0(x)}{dx} - \frac{d\log p_f(x)}{dx}\right)^2 dx$$

$$= \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_0(x)}{dx}\right)^2 dx + \frac{1}{2}\int_a^b p_0(x)\left(\frac{d\log p_f(x)}{dx}\right)^2 dx$$

$$- \int_a^b p_0(x)\left(\frac{d\log p_f(x)}{dx}\right)\left(\frac{d\log p_0(x)}{dx}\right) dx$$

Final term:

$$\int_a^b p_0(x)\left(\frac{d\log p_f(x)}{dx}\right)\left(\frac{d\log p_0(x)}{dx}\right) dx$$

$$= \int_a^b \cancel{p_0(x)}\left(\frac{d\log p_f(x)}{dx}\right)\left(\frac{1}{\cancel{p_0(x)}}\frac{dp_0(x)}{dx}\right) dx$$

$$= \left[\left(\frac{d\log p_f(x)}{dx}\right)p_0(x)\right]_a^b - \int_a^b p_0(x)\frac{d^2\log p_f(x)}{dx^2}.$$

Score matching (general version)

- Score matching (Hyvärinen, 2005) since MLE can be intractable even in finite dimensions if $A(\theta)$ is not easily computable.
- Assuming p_f to be differentiable (w.r.t. x) and $\int p_0(x)\|\nabla_x \log p_f(x)\|^2 \, dx < \infty$, $\forall\, \theta \in \Theta$:

$$
\begin{aligned}
D_F(p_0 \| p_f) \quad &:= \quad \frac{1}{2} \int p_0(x) \left\| \nabla_x \log p_0(x) - \nabla_x \log p_f(x) \right\|^2 \, dx \\
&\overset{(a)}{=} \quad \int p_0(x) \sum_{i=1}^{d} \left(\frac{1}{2} \left(\frac{\partial \log p_f(x)}{\partial x_i} \right)^2 + \frac{\partial^2 \log p_f(x)}{\partial x_i^2} \right) dx \\
&\qquad + \frac{1}{2} \int p_0(x) \left\| \frac{\partial \log p_0(x)}{\partial x} \right\|^2 dx,
\end{aligned}
$$

where partial integration is used in (a) under the condition that

$$
p_0(x) \frac{\partial \log p_f(x)}{\partial x_i} \to 0 \text{ as } x_i \to \pm\infty, \; \forall\, i = 1, \ldots, d.
$$

Empirical Estimator

p_n represents n i.i.d. samples from P_0

Empirical Estimator

p_n represents n i.i.d. samples from P_0

$$D_F(p_n \| p_f) := \frac{1}{n} \sum_{a=1}^{n} \sum_{i=1}^{d} \left(\frac{1}{2} \left(\frac{\partial \log p_f(X_a)}{\partial x_i} \right)^2 + \frac{\partial^2 \log p_f(X_a)}{\partial x_i^2} \right) + C$$

Since $D_F(p_n, p_f)$ is independent of $A(f)$,

$$f_n^* = \arg\min_{f \in \mathcal{F}} D_F(p_n, p_f)$$

should be easily computable, unlike the MLE.

Empirical Estimator

p_n represents n i.i.d. samples from P_0

$$D_F(p_n \| p_f) := \frac{1}{n}\sum_{a=1}^{n}\sum_{i=1}^{d}\left(\frac{1}{2}\left(\frac{\partial \log p_f(X_a)}{\partial x_i}\right)^2 + \frac{\partial^2 \log p_f(X_a)}{\partial x_i^2}\right) + C$$

Since $D_F(p_n, p_f)$ is independent of $A(f)$,

$$f_n^* = \arg\min_{f \in \mathcal{F}} D_F(p_n, p_f)$$

should be easily computable, unlike the MLE.

Add extra term $\lambda \|f\|_{\mathcal{H}}^2$ to regularize

How do we get a comptable solution?

$$p_f(x) = e^{\langle f, \phi(x) \rangle_{\mathcal{H}} - A(f)} q_0(x)$$

Thus

$$\frac{\partial}{\partial x} \log p_f(x) = \frac{\partial}{\partial x} \langle f, \phi(x) \rangle_{\mathcal{H}} + \frac{\partial}{\partial x} \log q_0(x).$$

How do we get a comptable solution?

$$p_f(x) = e^{\langle f, \phi(x) \rangle_{\mathcal{H}} - A(f)} q_0(x)$$

Thus

$$\frac{\partial}{\partial x} \log p_f(x) = \frac{\partial}{\partial x} \langle f, \phi(x) \rangle_{\mathcal{H}} + \frac{\partial}{\partial x} \log q_0(x).$$

Kernel trick for derivatives:

$$\frac{\partial}{\partial x_i} f(X) = \left\langle f, \frac{\partial}{\partial x_i} \phi(X) \right\rangle_{\mathcal{H}}$$

How do we get a comptable solution?

$$p_f(x) = e^{\langle f, \phi(x) \rangle_{\mathcal{H}} - A(f)} q_0(x)$$

Thus

$$\frac{\partial}{\partial x} \log p_f(x) = \frac{\partial}{\partial x} \langle f, \phi(x) \rangle_{\mathcal{H}} + \frac{\partial}{\partial x} \log q_0(x).$$

Kernel trick for derivatives:

$$\frac{\partial}{\partial x_i} f(X) = \left\langle f, \frac{\partial}{\partial x_i} \phi(X) \right\rangle_{\mathcal{H}}$$

Dot product between feature derivatives:

$$\left\langle \frac{\partial}{\partial x_i} \phi(X), \frac{\partial}{\partial x_j} \phi(X') \right\rangle_{\mathcal{H}} = \frac{\partial^2}{\partial x_i \partial x_{d+j}} k(X, X')$$

How do we get a comptable solution?

$$p_f(x) = e^{\langle f, \phi(x) \rangle_{\mathcal{H}} - A(f)} q_0(x)$$

Thus

$$\frac{\partial}{\partial x} \log p_f(x) = \frac{\partial}{\partial x} \langle f, \phi(x) \rangle_{\mathcal{H}} + \frac{\partial}{\partial x} \log q_0(x).$$

Kernel trick for derivatives:

$$\frac{\partial}{\partial x_i} f(X) = \left\langle f, \frac{\partial}{\partial x_i} \phi(X) \right\rangle_{\mathcal{H}}$$

Dot product between feature derivatives:

$$\left\langle \frac{\partial}{\partial x_i} \phi(X), \frac{\partial}{\partial x_j} \phi(X') \right\rangle_{\mathcal{H}} = \frac{\partial^2}{\partial x_i \partial x_{d+j}} k(X, X')$$

By representer theorem:

$$f_n^* = \alpha \hat{\xi} + \sum_{\ell=1}^{n} \sum_{j=1}^{d} \beta_{\ell j} \frac{\partial \phi(X_\ell)}{\partial x_j},$$

Consistency and Rates for $f_{\lambda,n}$

Suppose existence assumptions hold.

(i) (Consistency) If $f_0 \in \overline{\mathcal{R}(C)}$, then

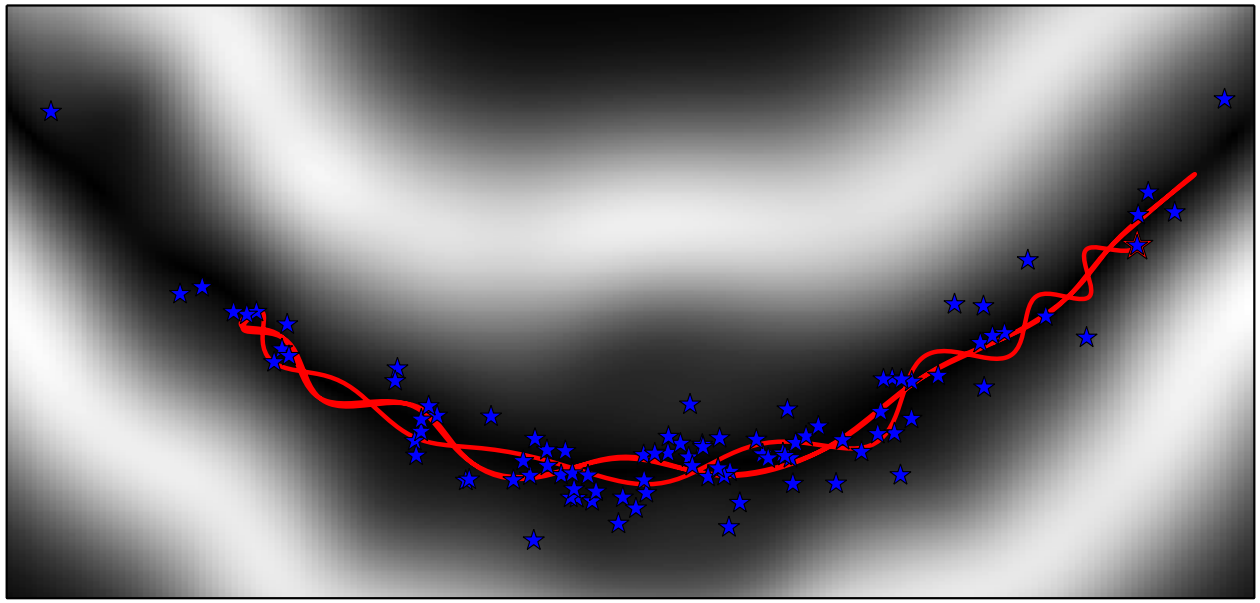
$$\|f_{\lambda,n} - f_0\|_{\mathcal{H}} \to 0 \text{ as } \lambda\sqrt{n} \to \infty,\ \lambda \to 0, \text{ and } n \to \infty.$$

(ii) (Rates of convergence) Suppose $f_0 \in \mathcal{R}(C^\beta)$ for some $\beta > 0$. Then

$$\|f_{\lambda,n} - f_0\|_{\mathcal{H}} = O_{p_0}\left(n^{-\min\left\{\frac{1}{4}, \frac{\beta}{2(\beta+1)}\right\}}\right)$$

with $\lambda = n^{-\max\left\{\frac{1}{4}, \frac{1}{2(\beta+1)}\right\}}$ as $n \to \infty$.

Kernel Adaptive Hamiltonian Monte Carlo (KMC)

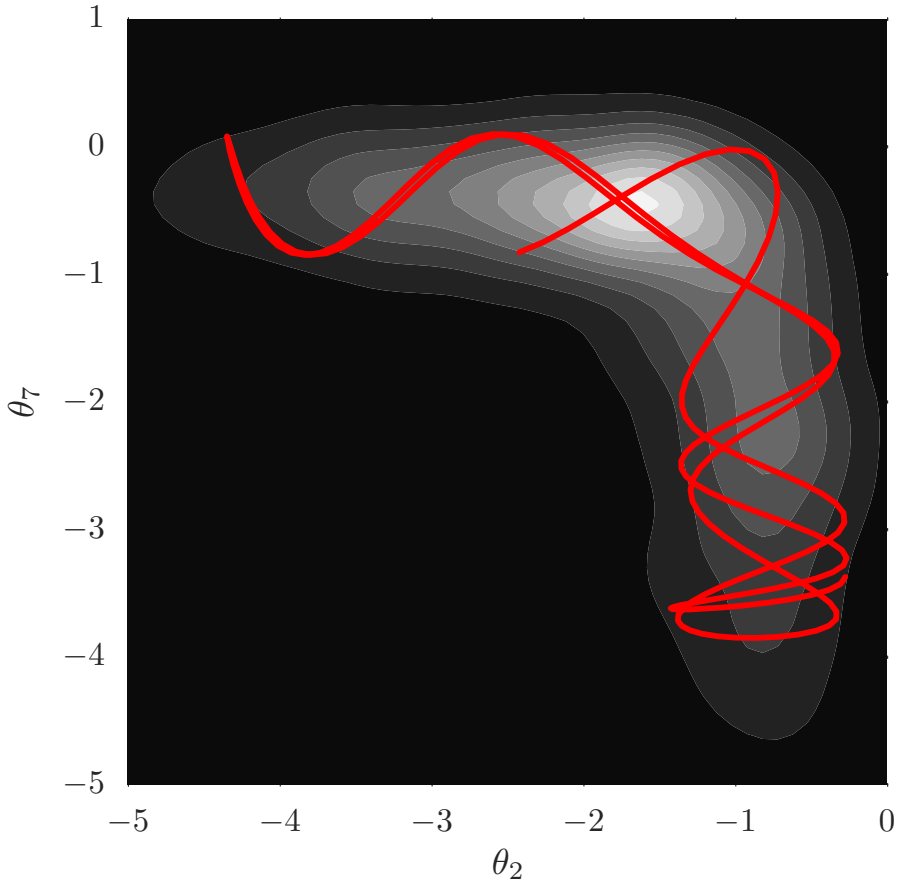


Heiko Strathmann, Dino Sejdinovic, Samuel Livingstone, Zoltan Szabo, and Arthur Gretton, NIPS 2015

Hamiltonian Monte Carlo

- HMC: distant moves, high acceptance probability.
- Potential energy $U(x) = -\log \pi(x)$, auxiliary momentum $p \sim \exp(-K(p))$, simulate for $t \in \mathbb{R}$ along Hamiltonian flow of $H(p, x) = K(p) + U(x)$, using operator

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



- Numerical simulation (i.e. leapfrog) depends on *gradient information*.

Infinite dimensional exponential families

Proposal is RKHS exponential family model [Fukumizu, 2009; Sriperumbudur et al. 2014], but accept using correct MH ratio (to correct for both model and leapfrog)

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

- 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: $f \in \mathcal{H}$.

Estimation of unnormalised density models from samples via score matching [Sriperumbudur et al. 2014]

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

Approximate solution: KMC lite

$$f(x) = \sum_{i=1}^{m} \alpha_i k(z_i, x)$$

- $\mathbf{z} \subseteq \mathbf{x}$ sub-sample, $m < n$.
- α from linear system

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

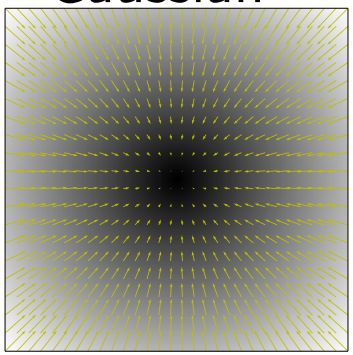
where $C \in \mathbb{R}^{m \times m}$, $b \in \mathbb{R}^m$ depend on kernel matrix

- Cost $\mathcal{O}(m^3 + m^2 d)$ (or cheaper with low-rank approx., conjugate gradient).

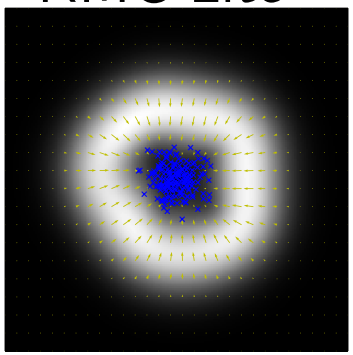
Geometrically ergodic on log-concave targets (fast convergence).

Gradient norm:

Gaussian



KMC Lite



Approximate solution: KMC finite

$$f(x) = \theta^\top \phi_x$$

- *Random Fourier Features* $\phi_x^\top \phi_y \approx k(x, y)$
- $\theta \in \mathbb{R}^m$ can be computed from

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

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

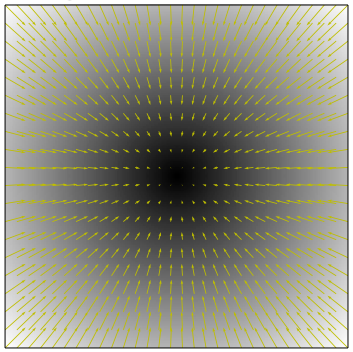
where $\dot{\phi}^{\ell}_x := \frac{\partial}{\partial x_\ell} \phi_x$ and $\ddot{\phi}^{\ell}_x := \frac{\partial^2}{\partial x_\ell^2} \phi_x$.

- *On-line updates* cost $\mathcal{O}(dm^2)$.

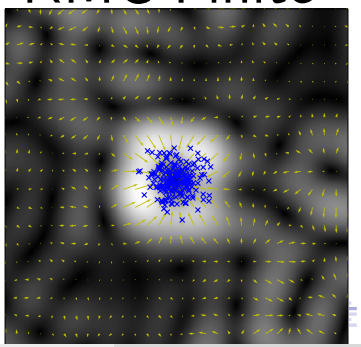
Updates fast, uses *all* Markov chain history. Caveat: need to initialise correctly.

Gradient norm:

Gaussian



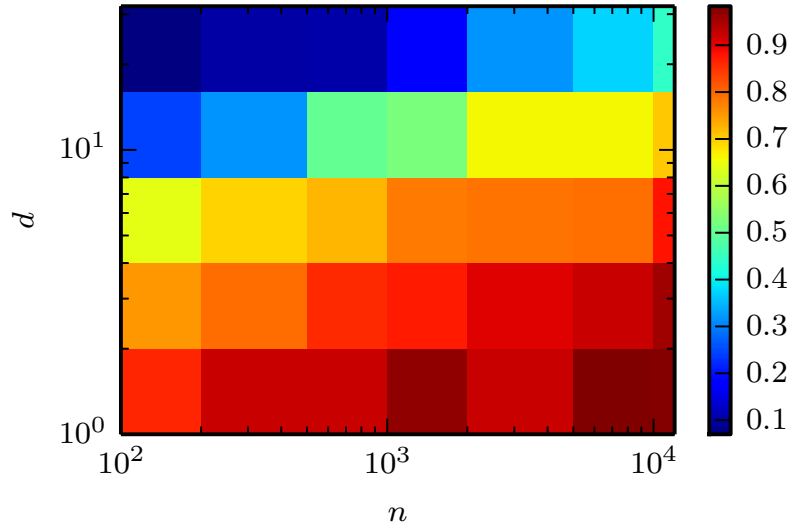
KMC Finite



Does kernel HMC work in high dimensions?

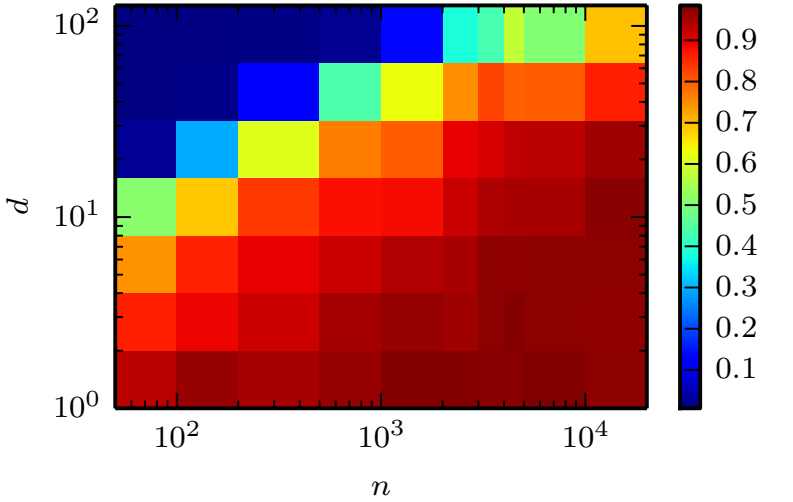
Challenging Gaussian target (**top**):

- Eigenvalues: $\lambda_i \sim \texttt{Exp}(1)$.
- 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 $d \approx 100$.



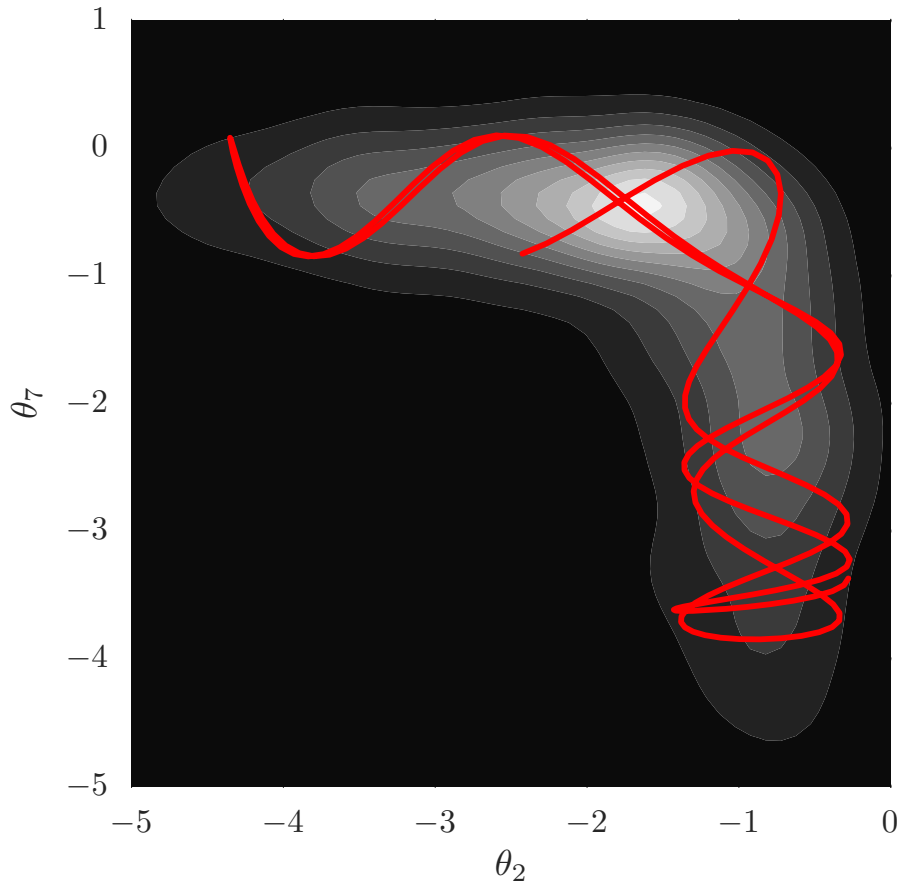
Gaussian Process Classification on UCI data

- 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.



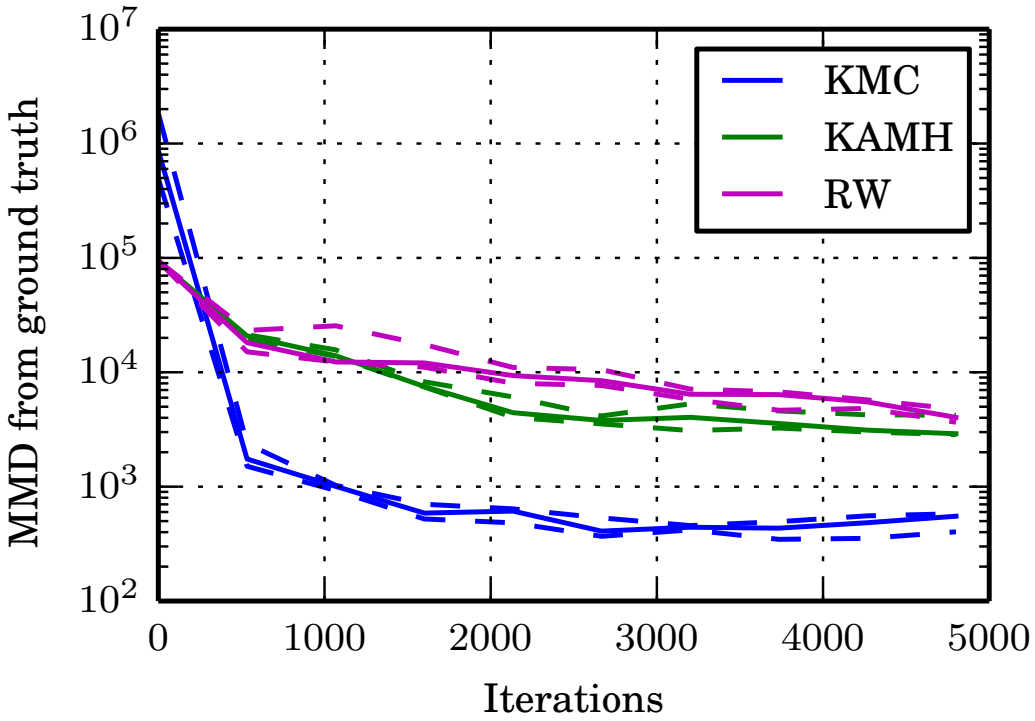
Gaussian Process Classification on UCI data

- 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.

Conclusions

- Kernel HMC: Simple, versatile, gradient-free adaptive MCMC sampler:
- Derivative of log density fit to samples, use this as proposal in HMC.
- Outperforms existing adaptive approaches on nonlinear target distributions
- Future work: how does convergence rate degrade with increasing dimension?

- **Kernel HMC code: https://github.com/karlnapf/kernel hmc**

Bayesian Gaussian Process Classification

Our case: target $\pi(\cdot)$ and log gradient not computable –
Pseudo-Marginal MCMC

Bayesian Gaussian Process Classification

Our case: target $\pi(\cdot)$ and log gradient not computable –
Pseudo-Marginal MCMC

Example: when is target not computable?

- GPC model: latent process $\mathbf{f}$, labels $\mathbf{y}$, (with covariate matrix X), and hyperparameters θ:

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

$\mathbf{f}|\theta \sim \mathcal{N}(0, \mathcal{K}_\theta)$ GP with covariance $\mathcal{K}_\theta$

Bayesian Gaussian Process Classification

Our case: target $\pi(\cdot)$ and log gradient not computable – **Pseudo-Marginal MCMC**

Example: when is target not computable?

- GPC model: latent process $\mathbf{f}$, labels $\mathbf{y}$, (with covariate matrix X), and hyperparameters θ:

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

  $\mathbf{f}|\theta \sim \mathcal{N}(0, \mathcal{K}_\theta)$ GP with covariance $\mathcal{K}_\theta$
- Automatic Relevance Determination (ARD) covariance:

$$(\mathcal{K}_\theta)_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}'_j|\theta) = \exp\left(-\frac{1}{2}\sum_{s=1}^{d}\frac{(x_{i,s} - x'_{j,s})^2}{\exp(\theta_s)}\right)$$

Bayesian Gaussian Process Classification

Our case: target $\pi(\cdot)$ and log gradient not computable – **Pseudo-Marginal MCMC**

Example: when is target not computable?

- GPC model: latent process $\mathbf{f}$, labels $\mathbf{y}$, (with covariate matrix X), and hyperparameters θ:

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

  $\mathbf{f}|\theta \sim \mathcal{N}(0, \mathcal{K}_\theta)$ GP with covariance $\mathcal{K}_\theta$
- Automatic Relevance Determination (ARD) covariance:

$$(\mathcal{K}_\theta)_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}'_j|\theta) = \exp\left(-\frac{1}{2}\sum_{s=1}^{d}\frac{(x_{i,s} - x'_{j,s})^2}{\exp(\theta_s)}\right)$$

- Classification $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f(x_i))$ where

$$p(y_i|f(x_i)) = (1 - \exp(-y_i f(x_i)))^{-1}, \qquad y_i \in \{-1, 1\}.$$

Pseudo-Marginal MCMC

Example: when is target not computable?

- Gaussian process classification, latent process $\mathbf{f}$

$$p(\theta|\mathbf{y}) \propto p(\theta)p(\mathbf{y}|\theta) = p(\theta)\int p(\mathbf{f}|\theta)p(\mathbf{y}|\mathbf{f},\theta)d\mathbf{f} =: \pi(\theta)$$

. . . but cannot integrate out $\mathbf{f}$

Pseudo-Marginal MCMC

Example: when is target not computable?

- Gaussian process classification, latent process $\mathbf{f}$

$$p(\theta|\mathbf{y}) \propto p(\theta)p(\mathbf{y}|\theta) = p(\theta)\int p(\mathbf{f}|\theta)p(\mathbf{y}|\mathbf{f},\theta)d\mathbf{f} =: \pi(\theta)$$

  ... but cannot integrate out $\mathbf{f}$

- MH ratio:

$$\alpha(\theta,\theta') = \min\left\{1, \frac{p(\theta')p(\mathbf{y}|\theta')q(\theta|\theta')}{p(\theta)p(\mathbf{y}|\theta)q(\theta'|\theta)}\right\}$$

Pseudo-Marginal MCMC

Example: when is target not computable?

- Gaussian process classification, latent process $\mathbf{f}$

$$p(\theta|\mathbf{y}) \propto p(\theta)p(\mathbf{y}|\theta) = p(\theta)\int p(\mathbf{f}|\theta)p(\mathbf{y}|\mathbf{f},\theta)d\mathbf{f} =: \pi(\theta)$$

  ... but cannot integrate out $\mathbf{f}$

- MH ratio:

$$\alpha(\theta,\theta') = \min\left\{1, \frac{p(\theta')p(\mathbf{y}|\theta')q(\theta|\theta')}{p(\theta)p(\mathbf{y}|\theta)q(\theta'|\theta)}\right\}$$

- **Filippone & Girolami, 2013** use Pseudo-Marginal MCMC: unbiased estimate of $p(\mathbf{y}|\theta)$ via importance sampling:

$$\hat{p}(\theta|\mathbf{y}) \propto p(\theta)\hat{p}(\mathbf{y}|\theta) \approx p(\theta)\frac{1}{n_{\text{imp}}}\sum_{i=1}^{n_{\text{imp}}} p(\mathbf{y}|\mathbf{f}^{(i)})\frac{p(\mathbf{f}^{(i)}|\theta)}{Q(\mathbf{f}^{(i)})}$$

Pseudo-Marginal MCMC

Example: when is target not computable?

- Gaussian process classification, latent process $\mathbf{f}$

$$p(\theta|\mathbf{y}) \propto p(\theta)p(\mathbf{y}|\theta) = p(\theta)\int p(\mathbf{f}|\theta)p(\mathbf{y}|\mathbf{f},\theta)d\mathbf{f} =: \pi(\theta)$$

  ... but cannot integrate out $\mathbf{f}$

- Estimated MH ratio:

$$\alpha(\theta,\theta') = \min\left\{1, \frac{p(\theta')\hat{p}(\mathbf{y}|\theta')q(\theta|\theta')}{p(\theta)\hat{p}(\mathbf{y}|\theta)q(\theta'|\theta)}\right\}$$

Pseudo-Marginal MCMC

Example: when is target not computable?

- Gaussian process classification, latent process $\mathbf{f}$

$$p(\theta|\mathbf{y}) \propto p(\theta)p(\mathbf{y}|\theta) = p(\theta)\int p(\mathbf{f}|\theta)p(\mathbf{y}|\mathbf{f},\theta)d\mathbf{f} =: \pi(\theta)$$

  ... but cannot integrate out $\mathbf{f}$

- Estimated MH ratio:

$$\alpha(\theta,\theta') = \min\left\{1, \frac{p(\theta')\hat{p}(\mathbf{y}|\theta')q(\theta|\theta')}{p(\theta)\hat{p}(\mathbf{y}|\theta)q(\theta'|\theta)}\right\}$$

- Replacing marginal likelihood $p(\mathbf{y}|\theta)$ with *unbiased estimate* $\hat{p}(\mathbf{y}|\theta)$ still results in correct invariant distribution **[Beaumont, 2003; Andrieu & Roberts, 2009]**

How rich is our density family?

Let $q_0 \in C_0(\Omega)$ be a probability density such that $q_0(x) > 0$ for all $x \in \Omega \subset \mathbb{R}^d$. Define

$$\mathcal{P}_c := \left\{ p \in C_0(\Omega) : \int_\Omega p(x)\, dx = 1, p(x) \geq 0, \forall\, x \in \Omega \text{ and } \left\| \frac{p}{q_0} \right\|_\infty < \infty \right\}.$$

How rich is our density family?

Let $q_0 \in C_0(\Omega)$ be a probability density such that $q_0(x) > 0$ for all $x \in \Omega \subset \mathbb{R}^d$. Define

$$\mathcal{P}_c := \left\{ p \in C_0(\Omega) : \int_\Omega p(x)\, dx = 1, p(x) \geq 0, \forall\, x \in \Omega \text{ and } \left\| \frac{p}{q_0} \right\|_\infty < \infty \right\}.$$

Suppose k satisfies

$(*)$ $k(\cdot, x) \in C_0(\Omega)$ for all $x \in \Omega \subset \mathbb{R}^d$;

$(**)$ $\int \int k(x,y) d\mu(x)\, d\mu(y) > 0$ for all $\mu \in M_b(\Omega) \backslash \{0\}$;

Then $\mathcal{P}$ is dense in $\mathcal{P}_c$ w.r.t. KL divergence, total variation and Hellinger distances.