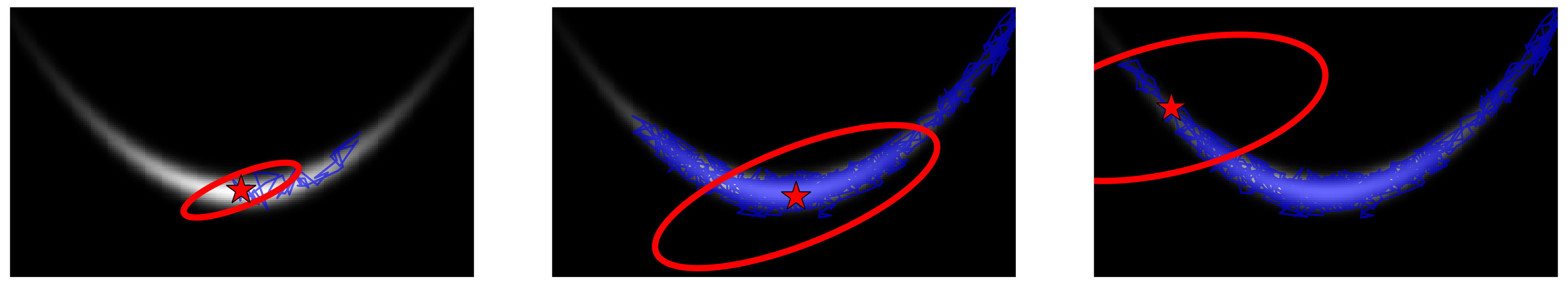
Kernel Adaptive Metropolis-Hastings

Dino Sejdinovic[1], Heiko Strathmann[1], Maria Lomeli Garcia[1], Christophe Andrieu[2], Arthur Gretton[1]

[1]Gatsby Unit, University College London. [2]School of Mathematics, University of Bristol

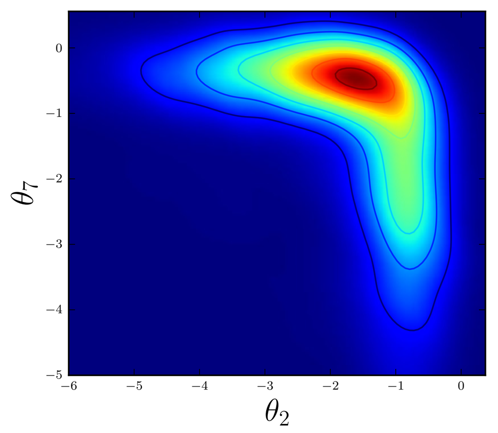
Adaptive Markov Chain Monte Carlo

- What proposal and scaling to choose for MCMC?
- Adaptive MCMC [1]: use history of Markov chain to learn structure of target, e.g. covariance.
- Only able to learn **global** linear covariance, i.e., scaling in principal directions.
- May be locally miscalibrated for strongly non-linear targets.



Motivation: Intractable & Non-linear Targets

- Non-linear targets: Hamiltonian Monte Carlo and MALA work great.
- However, those depend on gradients and second order information.
- Sometimes unavailable or expensive, e.g. in Bayesian GP classification, and more generally in Pseudo-Marginal MCMC [2].
- Right: Sliced posterior over hyperparameters of a GP classifier on UCI Glass.

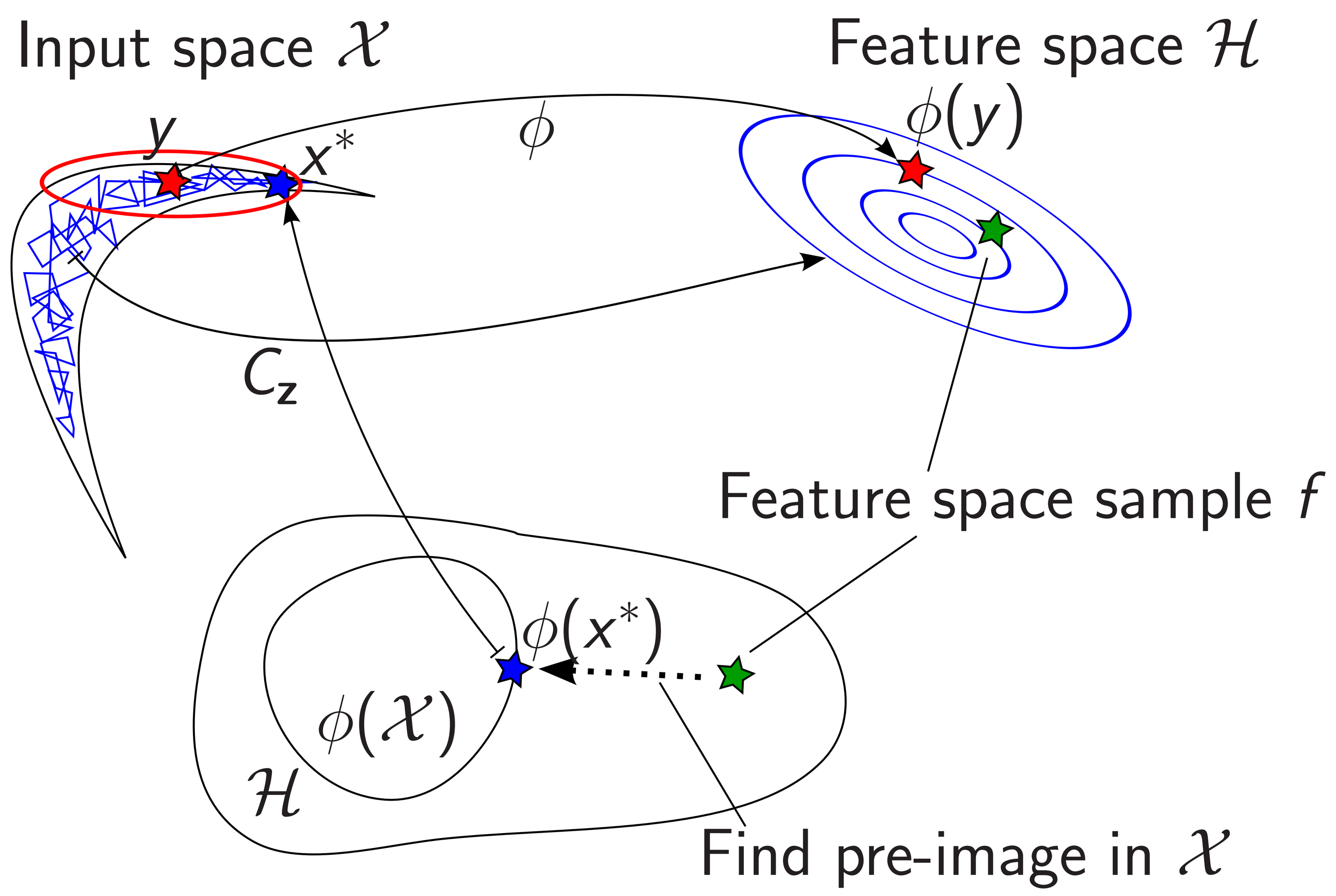


Want adaptive sampler that learns the shape of non-linear targets without higher order information.

Kernel Adaptive Metropolis Hastings – Illustration

Current point: y, a subsample of Markov chain history $\mathbf{z} = \{z_i\}_{i=1}^n$. Goal is an intelligent proposal x^*

- Capture non-linearities using linear covariance $C_\mathbf{z}$ in feature space $\mathcal{H}$.
- Sample $f \in \mathcal{H}$ from the Gaussian measure corresponding to $C_\mathbf{z}$.
- Find a point x^* whose feature mapping $\phi(x^*)$ is close to f



Embeddings and Covariance in RKHS

- For any positive semidefinite function k, there is a unique RKHS $\mathcal{H}_k$. Can consider $x \mapsto k(\cdot, x)$ as feature map.
- Embedding of a probability measure: $\mu_P = \int k(\cdot, x)\, dP(x)$ satisfies $\langle f, \mu_P \rangle_{\mathcal{H}_k} = \int f(x) dP(x) \quad \forall f \in \mathcal{H}_k$.
- Covariance operator: $C_P : \mathcal{H}_k \to \mathcal{H}_k$ is given by $C_P = \int k(\cdot, x) \otimes k(\cdot, x)\, dP(x) - \mu_P \otimes \mu_P$ [4]
- These can be estimated as

$$\mu_\mathbf{z} = \frac{1}{n}\sum_{i=1}^n k(\cdot, z_i) \qquad C_\mathbf{z} = \frac{1}{n}\sum_{i=1}^n k(\cdot, z_i) \otimes k(\cdot, z_i) - \mu_\mathbf{z} \otimes \mu_\mathbf{z}$$

The empirical covariance captures non-linear features of the underlying distribution (c.f. Kernel PCA [6])

Kernel Adaptive Metropolis Hastings – Formal Description

Current point: y, a subsample of Markov chain history $\mathbf{z} = \{z_i\}_{i=1}^n$. Goal is intelligent proposal x^*

1. Sample Gaussian Measure in RKHS: For $\beta \sim \mathcal{N}(0, \frac{\nu^2}{n} I_n)$, the represented RKHS element

$$f = k(\cdot, y) + \sum_{i=1}^n \beta_i [k(\cdot, z_i) - \mu_\mathbf{z}]$$

has mean $k(\cdot, y)$ and covariance $\frac{\nu^2}{n} C_\mathbf{z}$.

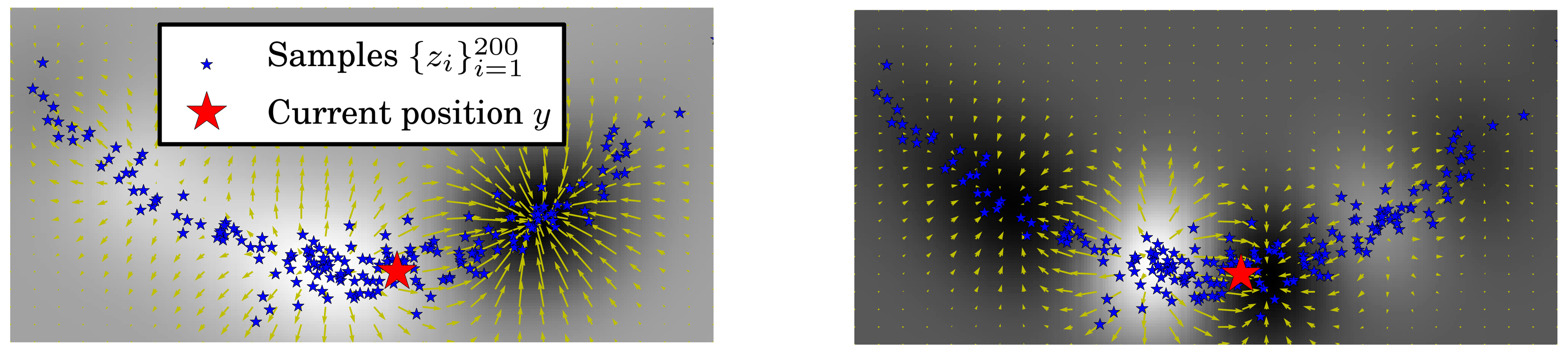
2. Find a point x^* in input space $\mathcal{X}$ with the feature embedding $\phi(x^*) = k(\cdot, x^*)$ close to f by considering

$$\arg\min_{x \in \mathcal{X}} \|k(\cdot, x) - f\|^2_{\mathcal{H}} = \arg\min_{x \in \mathcal{X}} \left\{ \underbrace{k(x,x) - 2k(x,y) - 2\sum_{i=1}^n \beta_i [k(x, z_i) - \mu_\mathbf{z}(x)]}_{=: g(x) \text{ where } g: \mathcal{X} \to \mathbb{R}} \right\},$$

taking a single gradient step w.r.t g, and (optionally) add 'exploration term' $\xi \sim \mathcal{N}(0, \gamma^2)$. This gives

$$x^* | y, \beta = y - \eta \nabla_x g(x)|_{x=y} + \xi = y - M_{\mathbf{z},y} H \beta + \xi,$$

where $M_{\mathbf{z},y} = 2\eta \left[\nabla_x k(x, z_1)|_{x=y}, \dots, \nabla_x k(x, z_n)|_{x=y}\right]$ is based on kernel gradients (readily available).



3. Integrating out RKHS samples and gradient step (i.e., β and ξ) gives Gaussian proposal on input space.

Proposed Algorithm: MCMC Kameleon

MCMC Kameleon

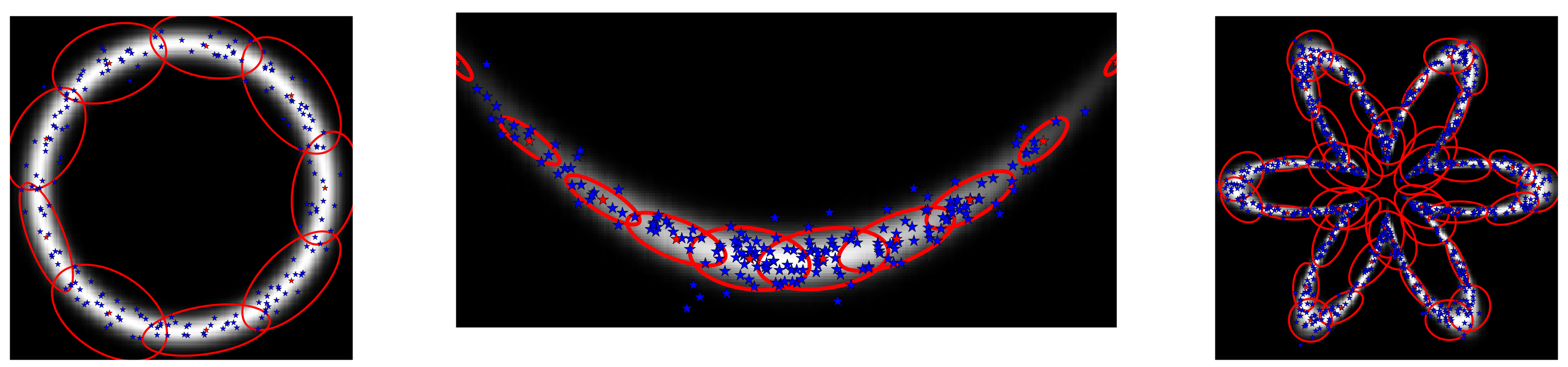
Input: unnormalized target π, subsample size n, scaling parameters ν, γ, kernel k,

At iteration $t+1$,

1. Obtain a random subsample $\mathbf{z} = \{z_i\}_{i=1}^n$ of the chain history $\{x_i\}_{i=0}^{t-1}$,
2. Sample proposed point x^* from $q_\mathbf{z}(\cdot|x_t) = \mathcal{N}(x_t, \gamma^2 I + \nu^2 M_{\mathbf{z},x_t} H M_{\mathbf{z},x_t}^\top)$,
3. Accept/Reject with standard MH ratio:

$$x_{t+1} = \begin{cases} x^*, & \text{w.p. } \min\left\{1, \frac{\pi(x^*) q_\mathbf{z}(x_t|x^*)}{\pi(x_t) q_\mathbf{z}(x^*|x_t)}\right\}, \\ x_t, & \text{otherwise.} \end{cases}$$

Straightforward to use in Pseudo-Marginal MCMC [2].



Kameleon proposals capture **local** covariance structure!

Examples of Covariance Structure for Standard Kernels

- **Linear kernel:** $k(x, x') = x^\top x'$

$$q_\mathbf{z}(\cdot|y) = \mathcal{N}(y, \gamma^2 I + 4\nu^2 \mathbf{Z}^\top H \mathbf{Z})$$

which results in the classical Adaptive Metropolis of [5].

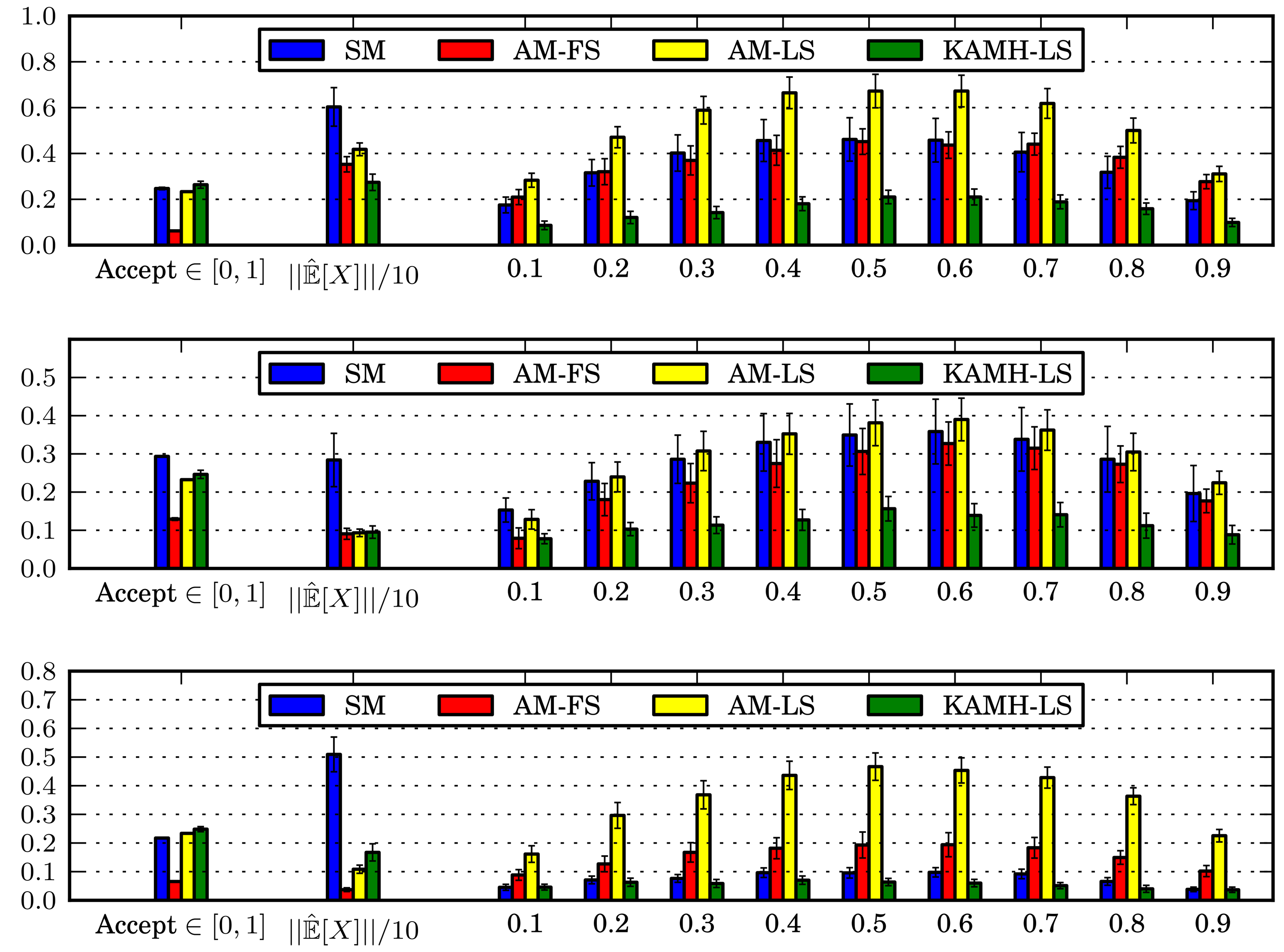
- **Gaussian kernel:** $k(x, x') = \exp\left(-\frac{1}{2}\sigma^{-2} \|x - x'\|_2^2\right)$

$$\left[\text{cov}[q_{\mathbf{z}(\cdot|y)}]\right]_{ij} = \gamma^2 \delta_{ij} + \frac{4\nu^2}{\sigma^4} \sum_{a=1}^n [k(y, z_a)]^2 (z_{a,i} - y_i)(z_{a,j} - y_j) + \mathcal{O}(n^{-1}),$$

where the previous points z_a influence the covariance, weighted by their similarity $k(y, z_a)$ to current point y.

Synthetic examples: Convergence Statistics

8-dim. Banana of [5]: moderately twisted (top), strongly twisted (middle) and 8-dim Flower (bottom).



Reported are acceptance rates and errors for means and quantiles.

Real-life Example: Bayesian Gaussian Process Classification

- Consider a standard GPC model

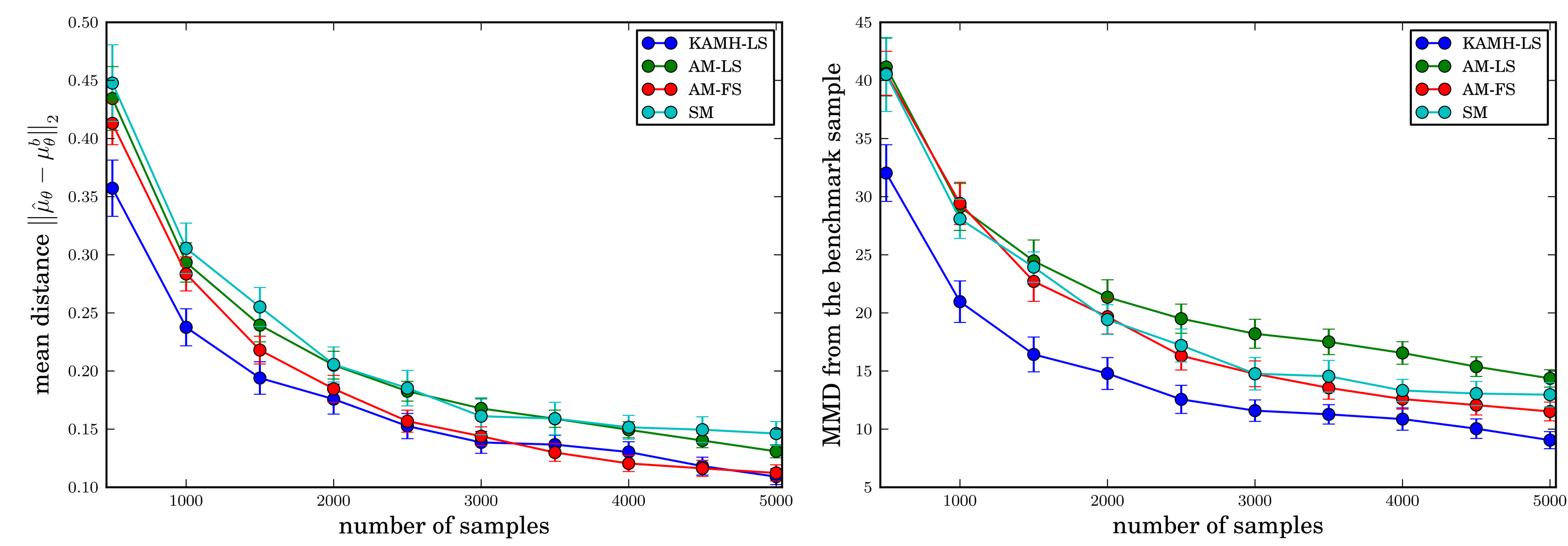
$$p(\mathbf{f}, \mathbf{y}, \theta) = p(\theta) p(\mathbf{f}|\theta) p(\mathbf{y}|\mathbf{f})$$

where $p(\mathbf{f}|\theta)$ is a Gaussian Process with an exponentiated quadratic covariance (ARD: one scale parameter per input space dimension), and $p(\mathbf{y}|\mathbf{f})$ is a sigmoidal function.

- Recent work [3] focused on Pseudo-Marginal MCMC to sample $p(\theta|y) = p(\theta) \int d\mathbf{f}\, p(\theta, \mathbf{f}|y) p(\mathbf{f}|\theta)$.
- Unbiased estimate of $\hat{p}(\mathbf{y}|\theta)$ via importance sampling with $q(\mathbf{f})$ obtained via Expectation Propagation:

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

- No access to likelihood, gradient, or Hessian of the target.



Performance on UCI Glass dataset: 8-dimensional non-linear posterior $p(\theta|\mathbf{y})$.

Literature

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

[2] Christophe Andrieu and Gareth O Roberts. The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

[3] Maurizio Filippone and Mark Girolami. Exact-approximate inference for Bayesian Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. To appear.

[4] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

[5] H. Haario, E. Saksman, and J. Tamminen. Adaptive Proposal Distribution for Random Walk Metropolis Algorithm. *Comput. Stat.*, 14(3):375–395, 1999.

[6] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10:1299–1319, 1998.

This work was funded by the Gatsby Charitable Foundation.

Code: `https://github.com/karlnapf/kameleon-mcmc`

Contact: {dino.sejdinovic, heiko.strathmann}@gmail.com