Kernel Adaptive Metropolis Hastings

Dino Sejdinovic, **Heiko Strathmann**, Maria Lomeli,
Christophe Andrieu, Arthur Gretton, ICML 2014

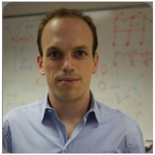
Gatsby Unit for Computational Neuroscience and Machine Learning
University College London

November 26, 2014

Joint work







Outline

Context: Intractable & non-linear Posteriors

Method: Kernel Embeddings & Covariance

Experiments: Results & Conclusion

Being Bayesian: Averaging beliefs of the unknown

$$p(y^*) = \int d\theta \, \underbrace{p(y^*|\theta)}_{\text{likelihood}} \, \underbrace{p(\theta|y)}_{\text{posterior}}$$

where $p(\theta|y) \propto p(y|\theta) \underbrace{p(\theta)}_{\text{prior}}$

Metropolis Hastings & Markov Chains

Construct $\theta_0 \to \theta_1 \to \theta_2 \to \theta_3 \to \dots$

- Given unnormalised target $\pi(\theta) \propto p(\theta|y)$
- At iteration t, state θ_t
- Propose $\theta' \sim q(\cdot|\theta_t)$

Accept $\theta_{t+1} \leftarrow \theta'$ with probability

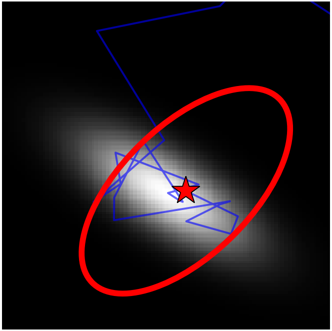
$$\min\left(\frac{\pi(\theta')q(\theta_t|\theta')}{\pi(\theta_t)q(\theta'|\theta_t)}, 1\right)$$

Reject $\theta_{t+1} \leftarrow \theta_t$ otherwise.

This talk: Which proposal?

Crucial for efficiency of sampler. Often,

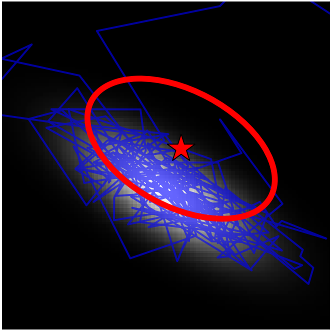
$$q_t(\cdot|\theta_t) = \mathcal{N}(\cdot|\theta_t, \dots)$$



Adaptive Metropolis: (Haario et al, 2001)

Online estimates of global covariance

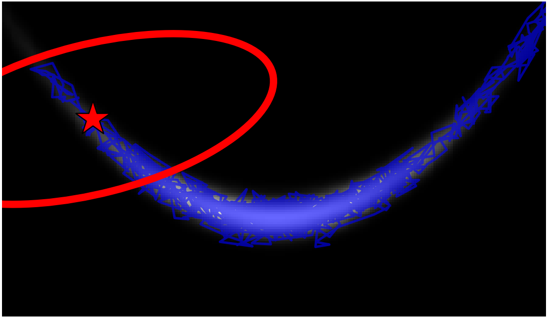
$$q_t(\cdot|\theta_t) = \mathcal{N}(\cdot|\theta_t, \nu^2 \hat{\Sigma}_t)$$



Adaptive Metropolis: (Haario et al, 2001)

Inefficient for curved targets

$$q_t(\cdot|\theta_t) = \mathcal{N}(\cdot|\theta_t, \nu^2\hat{\Sigma}_t)$$



Non-linear & Intractable Targets

Sophisticated solutions for non-linear targets:

- Metropolis Adjusted Langevin Algorithms (MALA), (Roberts & Stramer, 2003)
- Hamiltonian Monte Carlo (HMC), (Girolami & Calderhead, 2011)
- Require target gradient $\nabla\pi(\cdot)$ or second order information

Our case: Neither $\nabla\pi(\cdot)$ nor even $\pi(\cdot)$ can be computed.

Pseudo Marginal MCMC

- Posterior inference over latent process $\mathbf{f}$

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

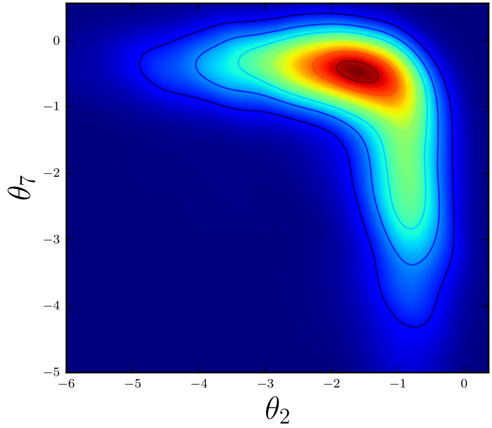
- Intractable for, e.g., non-conjugate Gaussian process
- Use unbiased estimator $\hat{\pi}(\cdot)$ in MH ratio
  Beaumont, 2003; Andrieu & Roberts, 2009; Filippone & Girolami 2014

$$\min\left(\frac{\hat{\pi}(\theta')q(\theta_t|\theta')}{\hat{\pi}(\theta_t)q(\theta'|\theta_t)}, 1\right)$$

- $\theta^{(j)}$ from correct invariant distribution
- No access to $\nabla\pi(\cdot)$

Gaussian Process Classification

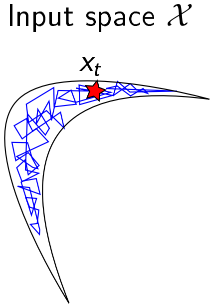
θ-Posterior slice of a GPC on UCI Glass dataset.



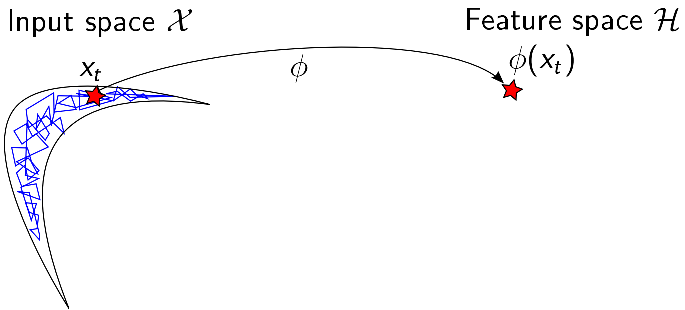
Objective: Adaptive sampler that learns the shape of non-linear targets without gradient information?

Method: Kernel Embeddings & Covariance

Proposal construction idea



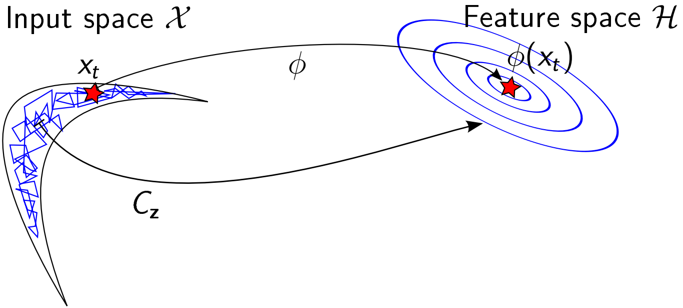
Proposal construction idea



Feature map & Kernel:

$$\phi : \mathcal{X} \to \mathcal{H} \qquad k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$$

Proposal construction idea

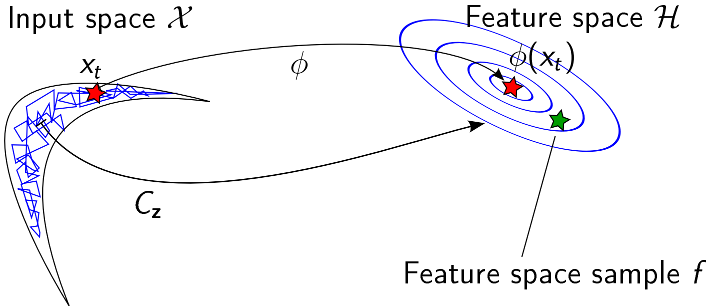


Kernel mean & covariance:

$$\mu := \mathbb{E}[\phi(X)] \qquad C := \mathbb{E}\left[\phi(X) \otimes \phi(X)\right] - \mu \otimes \mu$$

$$\hat{\mu}_\mathbf{z} = \frac{1}{n}\sum_{i=1}^{n} \phi(z_i) \qquad C_\mathbf{z} = \frac{1}{n}\sum_{i=1}^{n} \phi(z_i) \otimes \phi(z_i) - \mu_\mathbf{z} \otimes \mu_\mathbf{z}$$

Proposal construction idea

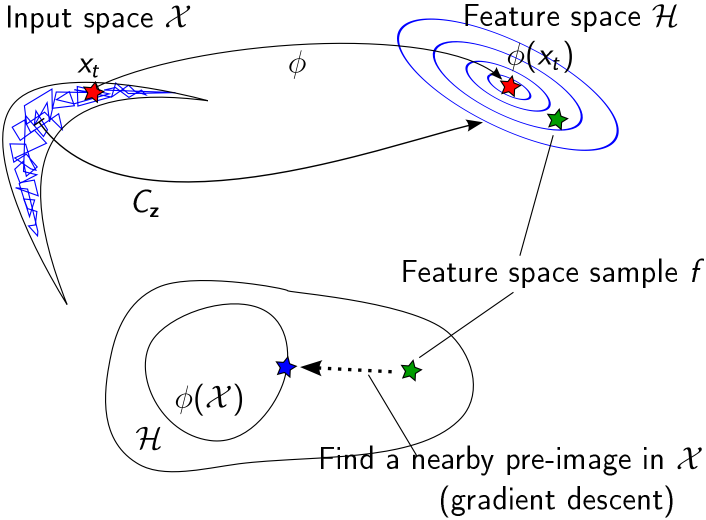


Feature space sample:

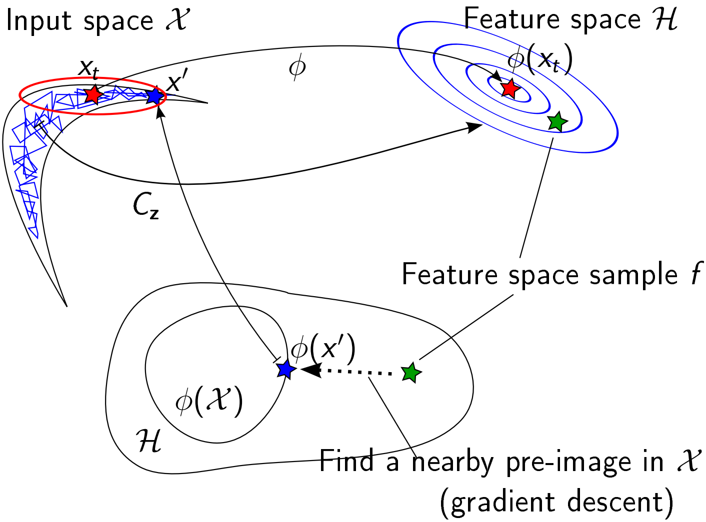
$$\beta \sim \mathcal{N}\left(\beta \middle| \mathbf{0}, \frac{\nu^2}{n} I_{n\times n}\right)$$

$$f = \phi(x_t) + \sum_{i=1}^{n} \beta_i \left[\phi(z_i) - \mu_\mathbf{z}\right]$$

Proposal construction idea



Proposal construction idea



Proposal construction formally

1. Get a chain subsample $\mathbf{z} = \{z_i\}_{i=1}^n$
2. Construct an RKHS sample $f \sim \mathcal{N}(\phi(x_t), \nu^2 C_{\mathbf{z}})$
3. Propose x' such that $\phi(x')$ is close to f i.e. attempt

$$x' = \arg\min_{x \in \mathcal{X}} \|\phi(x) - f\|_{\mathcal{H}}^2$$

4. Add noisy exploration term $\xi \sim \mathcal{N}(0, \gamma^2)$

This gives

$$x' | x_t, f, \xi = x_t - \eta \nabla_{x=x_t} \|\phi(x) - f\|_{\mathcal{H}}^2 + \xi$$

Final proposal

We have

$$x'|x_t, f, \xi = x_t - \eta \nabla_{x=x_t} \|\phi(x) - f\|_{\mathcal{H}}^2 + \xi$$

Analytically integrate out

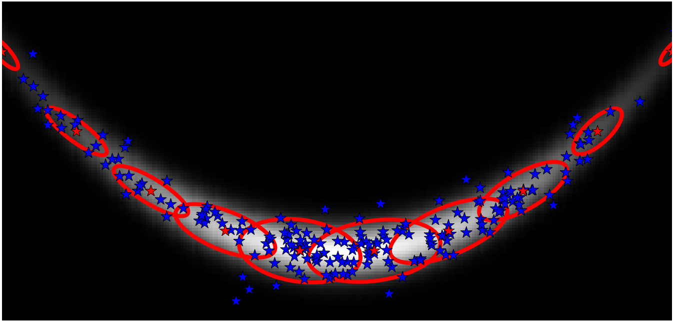
- RKHS samples f
- gradient step
- exploration noise ξ

Obtain Gaussian proposal on the input space:

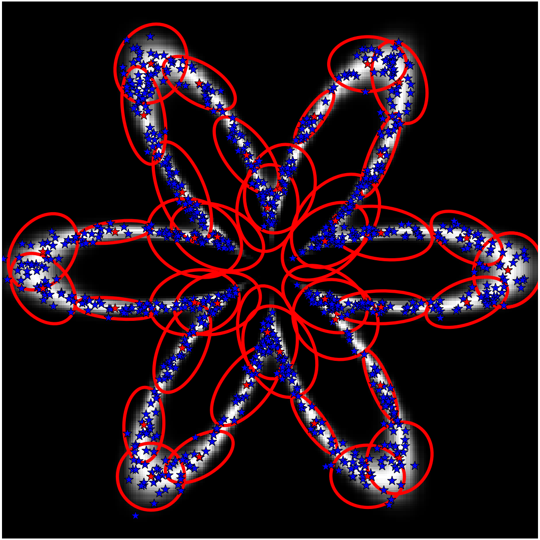
$$q_{\mathbf{z}}(x'|x_t) = \mathcal{N}(x_t, \gamma^2 I_d + \nu^2 M_{\mathbf{z},x_t} H M_{\mathbf{z},x_t}^\top)$$

$$M_{\mathbf{z},x_t} = 2\left[\nabla_{x=x_t} k(x, z_1), \ldots, \nabla_{x=x_t} k(x, z_n)\right]$$

Locally aligned covariance



Locally aligned covariance



Covariance structure for standard kernels

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

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

Classical Adaptive Metropolis (**Haario et al 1999;2001**)

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

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

Influence of previous points z_a on covariance is weighted by similarity $k(y, z_a)$ to current location y.

MCMC Kameleon

Input:

- unnormalized target π, or even $\hat{\pi}$
- kernel k
- subsample size n
- scaling parameters ν, γ
- update schedule $\{p_t\}_{t\geq 1}$ with $p_t \to 0$, $\sum_{t=1}^{\infty} p_t = \infty$

MCMC Kameleon

At iteration $t+1$,

1. With probability p_t, update 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 $q_{\mathbf{z}}(\cdot|x_t) = \mathcal{N}(x_t, \gamma^2 I_d + \nu^2 M_{\mathbf{z},x_t} H M_{\mathbf{z},x_t}^\top)$
3. Accept/reject with MH ratio

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

Convergence to π preserved as long as $p_t \to 0$ (Roberts & Rosenthal, 2007)

Outline

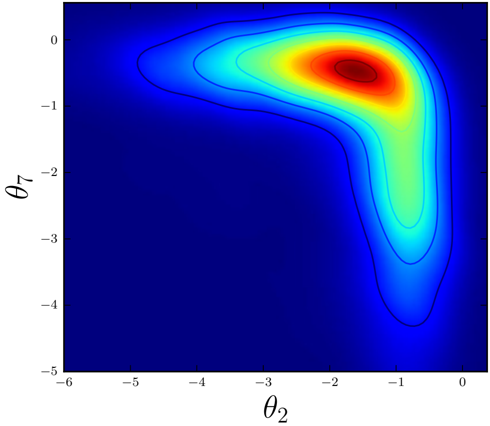
Context: Intractable & non-linear Posteriors

Method: Kernel Embeddings & Covariance

Experiments: Results & Conclusion

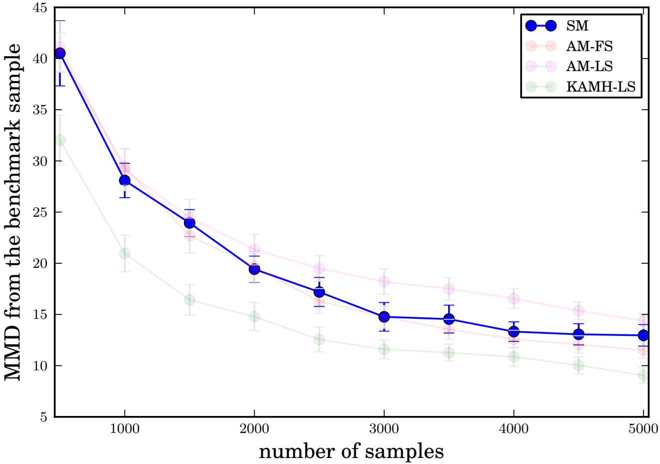
Gaussian Process Classification

Posterior over parameters of a GPC on UCI Glass dataset.



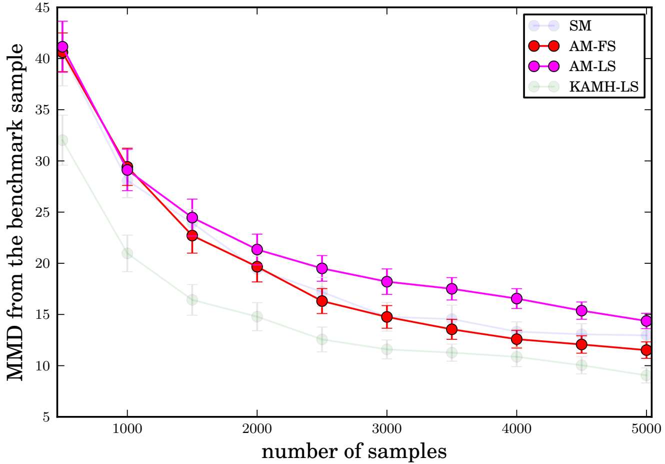
8-dimensional non-linear posterior: no ground truth, performance with respect to a long-run, heavily thinned benchmark sample. Mixed (1st-3rd) moments convergence.

UCI Glass dataset



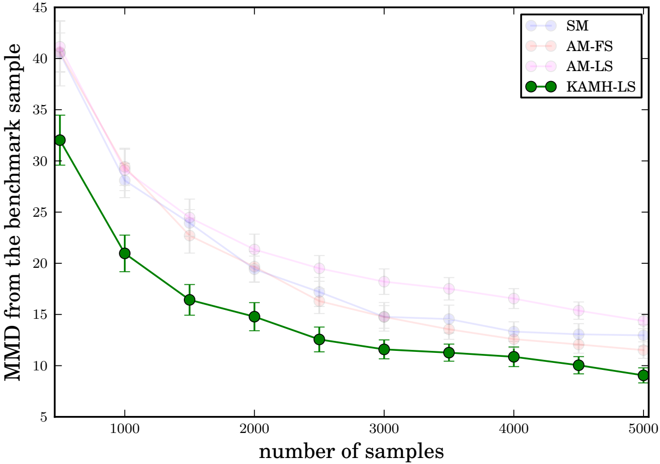
8-dimensional non-linear posterior: no ground truth, performance with respect to a long-run, heavily thinned benchmark sample. Mixed (1st-3rd) moments convergence.

UCI Glass dataset



8-dimensional non-linear posterior: no ground truth, performance with respect to a long-run, heavily thinned benchmark sample. Mixed (1st-3rd) moments convergence.

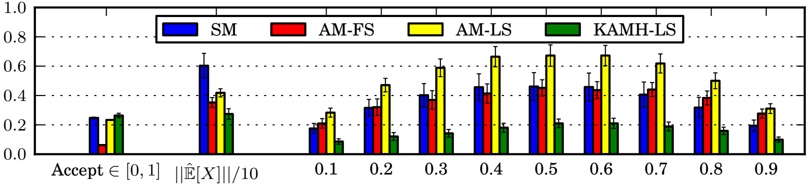
UCI Glass dataset



8-dimensional non-linear posterior: no ground truth, performance with respect to a long-run, heavily thinned benchmark sample. Mixed (1st-3rd) moments convergence.

Synthetic target: Banana

Banana: $\mathcal{B}(b, v)$: take $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \text{diag}(v, 1, \ldots, 1)$, and set $Y_2 = X_2 + b(X_1^2 - v)$, and $Y_i = X_i$ for $i \neq 2$



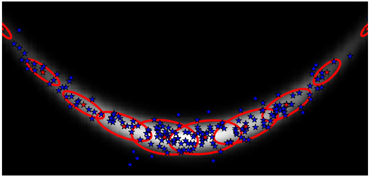
Strongly twisted 8-dimensional $\mathcal{B}(\mathbf{0.1}, \mathbf{100})$

Conclusions

MCMC Kameleon

- A simple, versatile, gradient-free adaptive MCMC sampler
- Proposals locally align with target distribution
- Outperforms existing approaches on nonlinear targets
- Very general framework (non-Euclidean $\mathcal{X}$)

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



Conclusions

MCMC Kameleon

- A simple, versatile, gradient-free adaptive MCMC sampler
- Proposals locally align with target distribution
- Outperforms existing approaches on nonlinear targets
- Very general framework (non-Euclidean $\mathcal{X}$)

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

Thank you! Questions?

RKHS and Kernel Embedding

Definition

Let k be a kernel on $\mathcal{X}$, and P a probability measure on $\mathcal{X}$. The kernel embedding of P into the RKHS $\mathcal{H}_k$ is $\mu_k(P) \in \mathcal{H}_k$ such that

$$\mathbb{E}_P f(X) = \langle f, \mu_k(P) \rangle_{\mathcal{H}_k}$$

for all $f \in \mathcal{H}_k$

- For any positive semidefinite function k, there is a unique RKHS $\mathcal{H}_k$. Can consider $x \mapsto k(\cdot, x)$ as a feature map.
- For many kernels k, including the Gaussian, Laplacian and inverse multi-quadratics, the kernel embedding $P \mapsto \mu_P$ is injective (Sriperumbudur et al, 2010)
- Captures all moments (similarly to the characteristic function).

Covariance Operator

Definition

The covariance operator of P is $C_P : \mathcal{H}_k \to \mathcal{H}_k$ such that $\forall f, g \in \mathcal{H}_k$,

$$\langle f, C_P g \rangle_{\mathcal{H}_k} = \mathrm{Cov}_P\left[f(X)g(X)\right]$$

.

$C_P : \mathcal{H}_k \to \mathcal{H}_k$ is given by

$$C_P = \int k(\cdot, x) \otimes k(\cdot, x)\, dP(x) - \mu_P \otimes \mu_P$$

(covariance of canonical features), and for $f, g, h \in \mathcal{H}_k$

$$\langle f \otimes g \rangle_{\mathcal{H}_k} h := \langle h, g \rangle_{\mathcal{H}_k} f$$

Feature space sample

RKHS sample

$$f = \phi(x_t) + \sum_{i=1}^{n} \beta_i \left[\phi(z_i) - \mu_{\mathbf{z}}\right]$$

has covariance

$$\begin{aligned} & \mathbb{E}\left[(f - \phi(x_t)) \otimes (f - \phi(x_t))\right] \\ = & \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \beta_i \beta_j (\phi(z_i) - \mu_{\mathbf{z}}) \otimes (\phi(z_j) - \mu_{\mathbf{z}})\right] \\ = & \frac{\nu^2}{n}\sum_{i=1}^{n}(\phi(z_i) - \mu_{\mathbf{z}}) \otimes (\phi(z_i) - \mu_{\mathbf{z}}) \\ = & \nu^2 C_{\mathbf{z}} \end{aligned}$$

Kernel distance gradient

$$
\begin{aligned}
g(x) &= \|\phi(x) - f\|^2_{\mathcal{H}} \\
&= k(x,x) - 2k(x,y) - 2\sum_{i=1}^{n} \beta_i \left[k(x,z_i) - \mu_{\mathbf{z}}(x)\right] \\
\nabla_x g(x)|_{x=y} &= \underbrace{\nabla_x k(x,x)|_{x=y} - 2\nabla_x k(x,y)|_{x=y}}_{=0} - M_{\mathbf{z},y} H \beta
\end{aligned}
$$

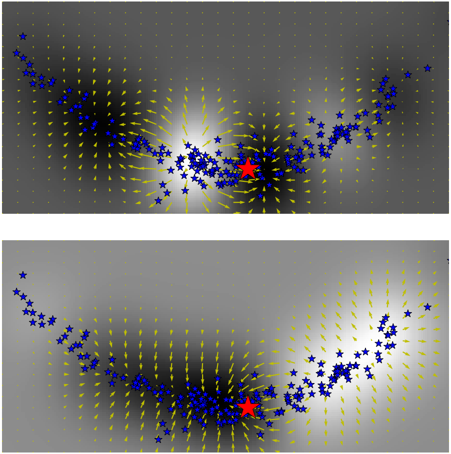
where

$$
M_{\mathbf{z},y} = 2\left[\nabla_x k(x,z_1)|_{x=y}, \ldots, \nabla_x k(x,z_n)|_{x=y}\right]
$$

and

$$
H = I_n - \frac{1}{n}\mathbf{1}_{n\times n}
$$

Gradient step intuition



$\|\phi(x) - f\|_{\mathcal{H}}^2$ varies most along high density areas of $\pi(\cdot)$

Bayesian Gaussian Process Classification

- GPC model: latent process $\mathbf{f}$, labels $\mathbf{y}$, covariates X, and hyperparameters θ:

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

where $\mathbf{f}|\theta \sim \mathcal{N}(0, \mathcal{K}_\theta)$ is a realization of a GP with covariance $\mathcal{K}_\theta$ (evaluated at X)

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

- $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i)$ is a product of sigmoidal functions:

$$p(y_i|f_i) = \frac{1}{1 - \exp(-y_i f_i)}, \qquad y_i \in \{-1, 1\}.$$

Bayesian Gaussian Process Classification

- Fully Bayesian treatment: Interested in posterior $p(\theta|y)$
- Cannot use a Gibbs sampler on $p(\theta, \mathbf{f}|y)$, which samples from $p(\mathbf{f}|\theta, y)$ and $p(\theta|\mathbf{f}, y)$ in turns, since $p(\theta|\mathbf{f}, y)$ is extremely sharp
- Filippone & Girolami, 2014 use Pseudo-Marginal MCMC to sample $p(\theta|y) = p(\theta) \int p(\theta, \mathbf{f}|y) p(\mathbf{f}|\theta) d\mathbf{f}$
- Unbiased estimate of $\hat{p}(\mathbf{y}|\theta)$ via importance sampling:

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

- No access to likelihood, gradient, or Hessian