Kernel Adaptive Metropolis-Hastings:

Gaussian Process Classification using Pseudo-marginal MCMC

Nicholas Sterge

April 22, 2018

Adaptive Metropolis-Hastings

Metropolis-Hastings:

- Propose $\theta^* \sim q(\cdot|\theta)$
- Accept θ^* w.p. $\min\left\{1, \frac{\pi(\theta^*)q(\theta|\theta^*)}{\pi(\theta)q(\theta^*|\theta)}\right\}$

What is the best proposal, $q(\cdot|\theta)$?

- Adaptive M-H: learn covariance structure of target, adapt proposal accordingly
- Gaussian w/ covariance learned from chain history, $\mathbf{q_t}(\cdot|\theta^{(\mathbf{t})}) = \mathcal{N}(\cdot|\theta^{(\mathbf{t})}, \nu^{\mathbf{2}}\mathbf{\Sigma_t})$ Haario et al. (1999)
- Adaptive scaling, principal component updates Andrieu and Thoms (2008)

Adaptive Metropolis-Hastings

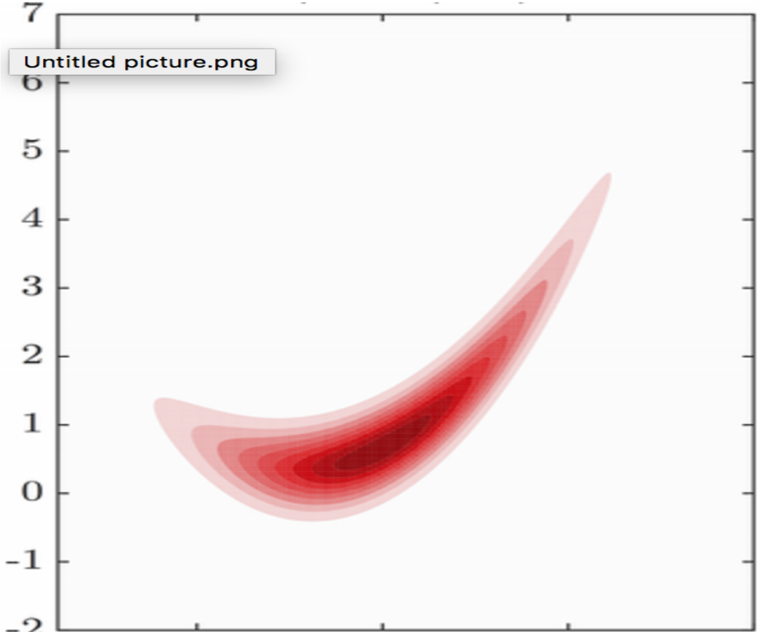
Why adapt?

- Limited information about target
- Multiple proposals to tune, e.g. Section 8 of Haario et al. (1999) (GOMOS)
- Adaptive burn-in for Metropolis algorithm

Shortcomings:

- Strongly non-linear targets, e.g. Banana target of Haario et al. (1999), Flower target of Sejdinovic et al. (2014)
- Directions of large variance depend on location of sampler

Banana



Kernel Adaptive Metropolis-Hastings

Motivation:

Principal Component Proposals

- Estimate $\Sigma_{\mathbf{z}}$ from subset of chain history, $\mathbf{z}$
- Eigenvectors/eigenvalues inform proposal, i.e. mixture of random walks down principal eigendirections

Kernelize $\rightarrow$ Kernel Principal Component Proposals

- Using kernel PCA, nonlinear principal directions can inform our proposal

Kernel Adaptive Metropolis-Hastings

Idea: Nonlinear support of target may be learned using Kernel PCA

- RKHS of functions, $\mathcal{H}$, $f : \mathcal{X} \to \mathbb{R}$, with RK $k(\cdot, \cdot)$
- Probability measure $\mathbb{P}$ on $\mathcal{X}$, covariance operator $C_{\mathbb{P}}$, empirical covariance operator $C_{\mathbf{z}}$ ($\mathbf{z} = \{z_i\}_{i=1}^n$)
- Kernel PCA is linear PCA on the covariance operator $C_{\mathbf{z}}$ [Schölkopf et al. (1998)]
- Principal eigendirections will be non-linear functions

How do we use Kernel PCA to construct a proposal?

1. Mixture of random walks down principal eigendirections
2. Consider the Gaussian measure on $\mathcal{H}$ induced by $C_{\mathbf{z}}$

Constructing the Proposal

Suppose current state y, subset of chain history $\mathbf{z} = \{z_i\}_{i=1}^n$

- Gaussian measure on $\mathcal{H}$:

$$\mathcal{N}(f|k(\cdot,y),\nu^2 C_{\mathbf{z}}) \propto \exp\left\{\frac{-1}{2\nu^2}\big(f-k(\cdot,y)\big)^T C_{\mathbf{z}}^{-1}\big(f-k(\cdot,y)\big)\right\}$$

- Samples have form $f = k(\cdot,y) + \sum_{i=1}^n \beta_i k(\cdot,z_i)$, where $\beta \sim \mathcal{N}(0, \frac{\nu^2}{n}\mathbf{I}_n)$

Moving from $\mathcal{H}$ to $\mathcal{X}$

- Obtain sample $f \in \mathcal{H}$, want $x \in \mathcal{X}$ s.t. $k(\cdot,x)$ 'near' f, i.e. $x = \arg\min \|k(\cdot,x) - f\|_{\mathcal{H}}^2$
- Minimization expensive $\rightarrow$ one iteration of gradient descent

Constructing the Proposal

Closed form:

$$q_{\mathbf{z}}(\cdot|\mathbf{y}) = \mathcal{N}(\cdot|y, \gamma^2 \mathbf{I}_n + \nu^2 \mathbf{M}_{\mathbf{z},y} \mathbf{H} \mathbf{M}_{\mathbf{z},y}^T)$$

where γ is a GD parameter, $\mathbf{H}$ centering matrix,
$\mathbf{M}_{\mathbf{z},y} = 2(\nabla_x k(x, z_1)|_{x=y}, \ldots)$

Update Schedule and Convergence:

- Proposal updated each time we update $\mathbf{z}$
- Adaptation probabilities $\{p_t\}_{t=1}^{\infty}$ s.t. $p_t \to 0$ and $\sum_t p_t = \infty$, guarantee convergence to correct target

Properties:

- Only requires evaluation of unnormalized target, locally adaptive in input space $\mathcal{X}$

Proposal Algorithm

At iteration $t+1$, current state is x_t

1. With probability p_t update random subsample, $\mathbf{z}$, of chain history
2. Sample proposed x^* from $q_{\mathbf{z}}(\cdot|x_t) = \mathcal{N}(\cdot|x_t, \gamma^2 \mathbf{I}_n + \nu^2 \mathbf{M}_{\mathbf{z},x_t} \mathbf{H} \mathbf{M}_{\mathbf{z},x_t}^T)$
3. Accept/Reject with M-H probability

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

$$x_{t+1} = x_t \quad w.p. 1 - \min\left\{1, \frac{\pi(x^*) q_{\mathbf{z}}(x_t|x^*)}{\pi(x_t) q_{\mathbf{z}}(x^*|x_t)}\right\}$$

Gaussian Process Classification

Inputs: $\mathbf{X} = \{x_1, ..., x_n\}$, $x_i \in \mathbb{R}^d$

Labels: $\mathbf{Y} = \{y_1, ..., y_n\}$, $y_i \in \{\pm 1\}$

Latent: $\mathbf{f} = \{f_1, ..., f_n\}$, $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, K(\theta))$

$$\theta = \{\sigma, \tau_1^2, ...\tau_d^2\}, \quad K(\theta)_{ij} = \sigma \exp\left(\frac{-1}{2}(x_i - x_j)^T A(x_i - x_j)\right)$$

$$A^{-1} = diag\left(\tau_1^2, ..., \tau_d^2\right)$$

$$p(y_i = 1|f_i) = \Phi(y_i f_i)$$

Gaussian Process Classification

Prediction:

$$p(y_* = 1|\mathbf{Y}, \theta) = \int p(y_* = 1|f_*)p(f_*|\mathbf{f}, \mathbf{Y})p(\mathbf{f}, \theta|\mathbf{Y})df_* d\mathbf{f} d\theta$$

MCMC to approximate integration w.r.t. $p(\mathbf{f}, \theta|\mathbf{Y})$

$$p(y_* = 1|\mathbf{Y}, \theta) \approx \frac{1}{N}\sum_{i=1}^{N} \int p(y_*|f_*)p(f_*|\mathbf{f}^{(i)}, \theta^{(i)})df_*$$

Pseudo-marginalization

Elliptical slice sampling to draw $\mathbf{f}$ from $p(\mathbf{f}|\mathbf{Y}, \theta)$

Pseudo-marginal MCMC to draw θ from $p(\theta|\mathbf{Y})$

$$\tilde{z} = \frac{\tilde{p}(\mathbf{Y}|\theta^*)p(\theta^*)q_{\mathbf{z}}(\theta|\theta^*)}{\tilde{p}(\mathbf{Y}|\theta)p(\theta)q_{\mathbf{z}}(\theta^*|\theta)}$$

where

$$\tilde{p}(\mathbf{Y}|\theta) = \frac{1}{N_{imp}} \sum_{i=1}^{N_{imp}} \frac{p(\mathbf{Y}|\mathbf{f}_i)p(\mathbf{f}_i|\theta)}{r(\mathbf{f}_i|\mathbf{Y}, \theta)}$$

$r(\cdot|\mathbf{Y}, \theta)$ is importance function obtained via Laplace Approximation

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

H. Haario, E. Saksman, and J. Tamminen. Adaptive proposal distribution for random walk metropolis. *Comput. Stat.*, 14(3): 375–395, 1999.

B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

Dino Sejdinovic, Heiko Strathmann, Maria Lomeli Garcia, Christopher Andrieu, and Arthur Gretton. Kernel adaptive metropolis-hastings. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 2014.