Inference in Deep Gaussian Processes Using Stochastic Gradient Hamiltonian Monte Carlo

Marton Havasi[1] José Miguel Hernández-Lobato[123] Juan José Murillo-Fuentes[4]

[1]Department of Engineering, University of Cambridge [2]Microsoft Research Cambridge [3]Alan Turing Institute [4]Department of Signal Theory and Communications, University of Sevilla

Gaussian Processes

A GP is a non-parametric model that assumes a jointly Gaussian distribution for any finite set of inputs. It defines a posterior distribution over functions $f : \mathbb{R}^D \to \mathbb{R}$.

- $\boldsymbol{x}, \boldsymbol{y}$ - inputs and outputs
- $\boldsymbol{z}, \boldsymbol{u}$ - inducing inputs and outputs
- $p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{I}\sigma^2)$ - likelihood function
- $[K_{\boldsymbol{ab}}]_{ij} = k(a_i, b_j)$ - covariance matrix

The joint probability density function can be written as

$$p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u}),$$

$$p(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu} = K_{\boldsymbol{xz}}K_{\boldsymbol{zz}}^{-1}\boldsymbol{u}$

$\boldsymbol{\Sigma} = K_{\boldsymbol{xx}} - K_{\boldsymbol{xz}}K_{\boldsymbol{zz}}^{-1}K_{\boldsymbol{xz}}^T.$

Variational Inference in Gaussian Processes

The goal is to find the posterior distribution $p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})$. Variational inference uses an approximation to the posterior

$$q(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})$$

$$q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{m}, \boldsymbol{S})$$

This choice of posterior allows for the marginalization of $\boldsymbol{u}$:

$$q(\boldsymbol{f}|\boldsymbol{m}, \boldsymbol{S}) = \int p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u} = \mathcal{N}(\boldsymbol{f}|\tilde{\mu}, \tilde{\Sigma})$$

where $\tilde{\mu} = K_{\boldsymbol{xz}}K_{\boldsymbol{zz}}^{-1}\boldsymbol{m}$,

$\tilde{\Sigma} = K_{\boldsymbol{xx}} - K_{\boldsymbol{xz}}K_{\boldsymbol{zz}}^{-1}(K_{\boldsymbol{zz}} - \boldsymbol{S})K_{\boldsymbol{zz}}^{-1}K_{\boldsymbol{xz}}^T.$

The parameters are optimized using the Expectation Lower Bound (ELBO):

$$\log p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}\big[\log p(\boldsymbol{y}|\boldsymbol{f})\big] - \mathrm{KL}\big[q(\boldsymbol{u})||p(\boldsymbol{u})\big].$$

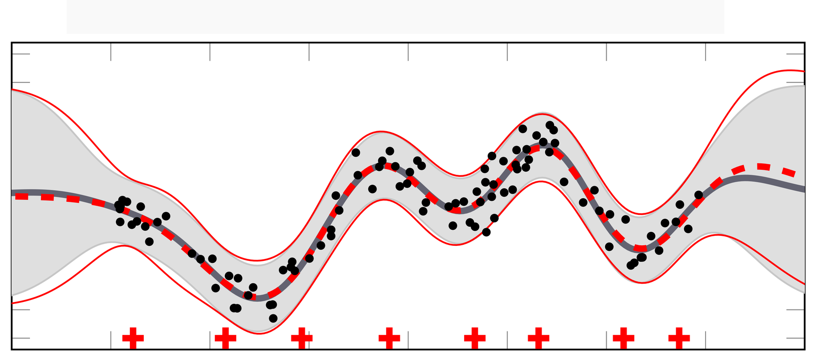


Figure: Sparse Gaussian Process
Source: Understanding Probabilistic Sparse Gaussian Process Approximations, NIPS 2016

Deep Gaussian Processes

Multilayer generalization of a Gaussian Processes. Unlike in neural networks, the uncertainty is propagated through the layers

$$p(\boldsymbol{y}, \{\boldsymbol{f}_l\}_{l=1}^L, \{\boldsymbol{u}_l\}_{l=1}^L) = p(\boldsymbol{y}|\boldsymbol{f}_L)\prod_{l=1}^L p(\boldsymbol{f}_l|\boldsymbol{u}_l)p(\boldsymbol{u}_l).$$

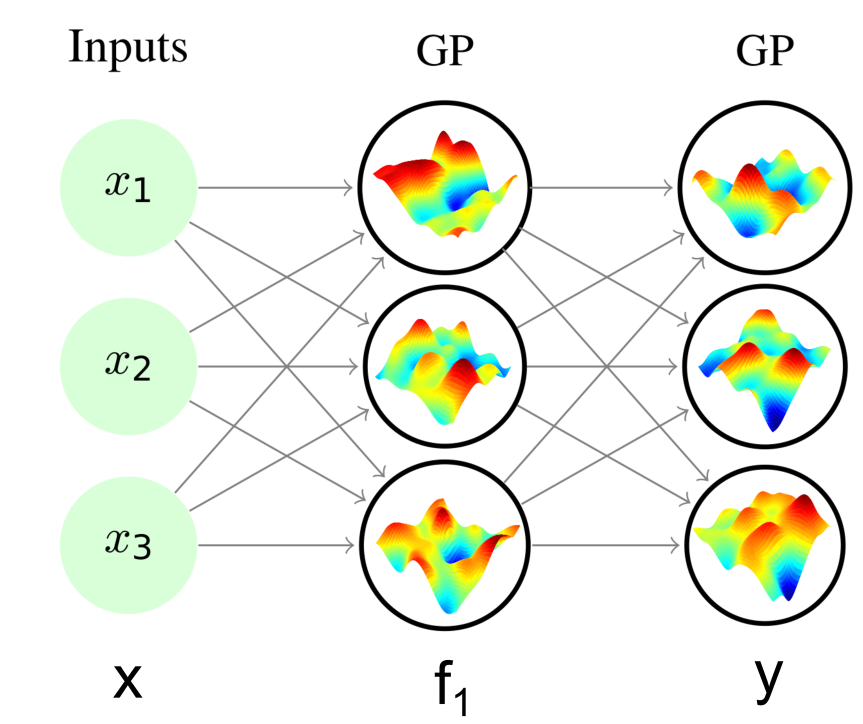


Figure: Deep Gaussian Process
Source: Avoiding pathologies in very deep networks, AIStats 2014

The ELBO takes a new form:

$$\log p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f}_L)}\big[\log p(\boldsymbol{y}|\boldsymbol{f}_L)\big] - \sum_{l=1}^L \mathrm{KL}\big[q(\boldsymbol{u}_l)||p(\boldsymbol{u}_l)\big].$$

Sampling from the Posterior

SGHMC is a MCMC sampling method that generates samples from the posterior purely from stochastic gradient estimates by mimicking Hamiltonian dynamics. $\boldsymbol{r}$ plays the role of the kinetic energy.

$$p(\boldsymbol{u}, \boldsymbol{r}|\boldsymbol{y}) \propto \exp\big(-U(\boldsymbol{u}) - \frac{1}{2}\boldsymbol{r}^T M^{-1}\boldsymbol{r}\big),$$

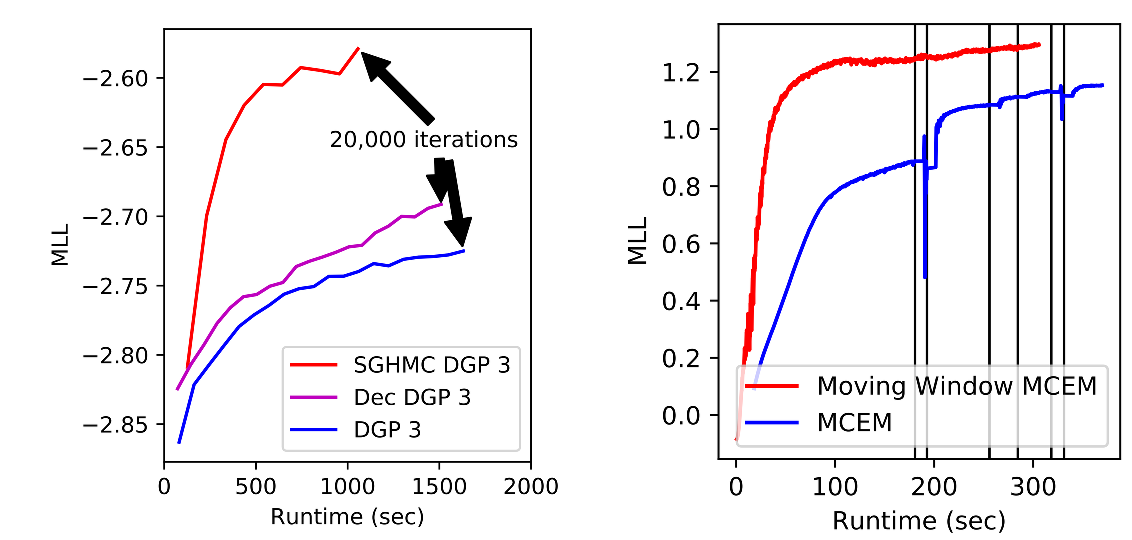
$$U(\boldsymbol{u}) = -\log p(\boldsymbol{u}|\boldsymbol{y}).$$

With update equations:

$$\Delta\boldsymbol{u} = \epsilon M^{-1}\boldsymbol{r},$$

$$\Delta\boldsymbol{r} = -\epsilon\nabla U(\boldsymbol{u}) - \epsilon C M^{-1}\boldsymbol{r} + \mathcal{N}\big(0, 2\epsilon(C - \hat{B})\big),$$

A further improvement, Moving Window MCEM recycles the recent posterior samples to speed up the training of the hyperparameters.



(a) Sampling is faster than Variational Inference (b) Moving Window MCEM versus MCEM

The posterior

A gaussian approximation may be insufficient. The posterior can be multimodal.

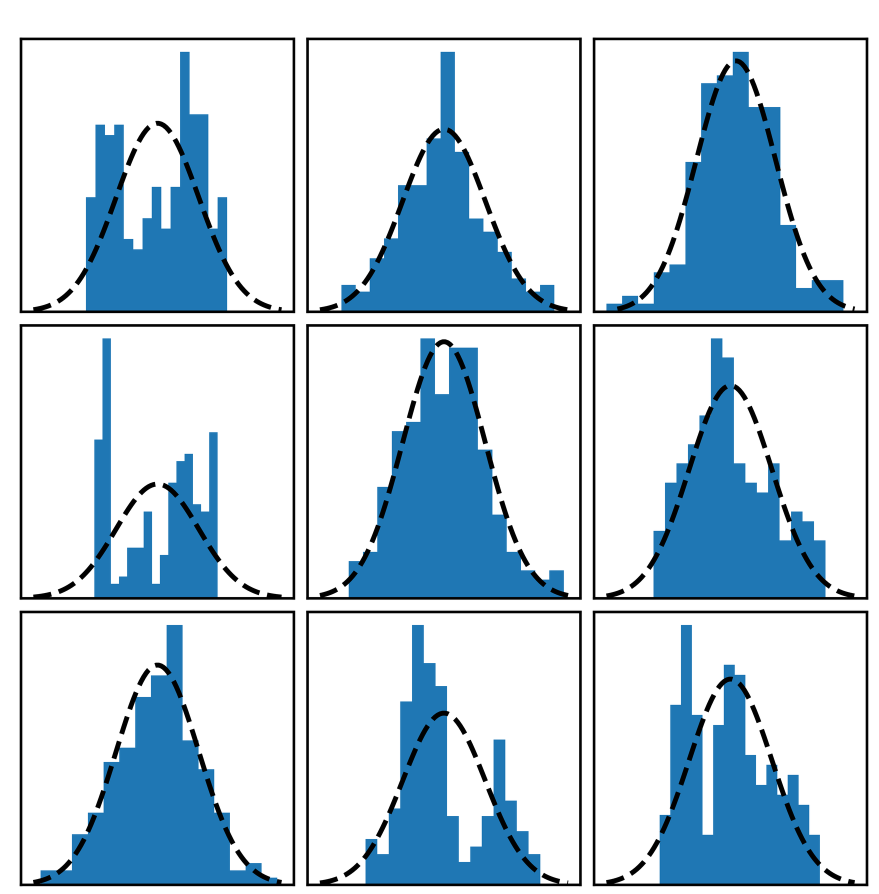


Figure: Samples from the posterior of $\boldsymbol{u}$.

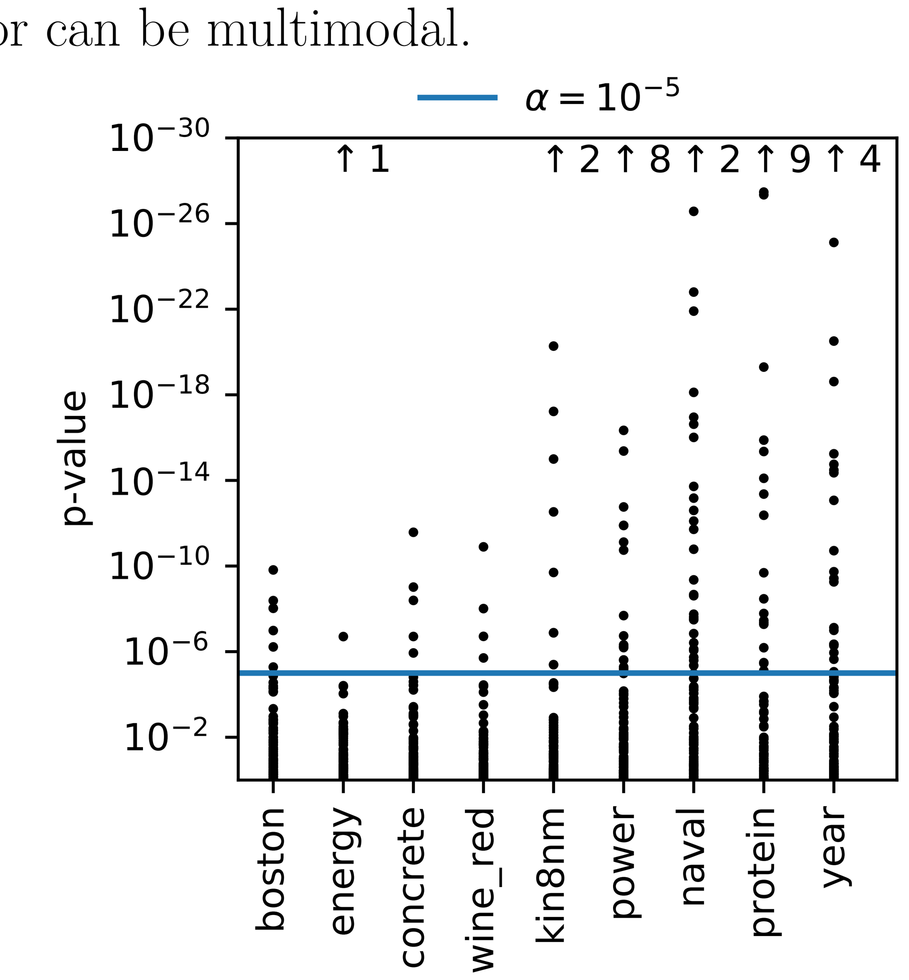


Figure: P-values corresponding to a Gaussian posterior.

Experimental Results

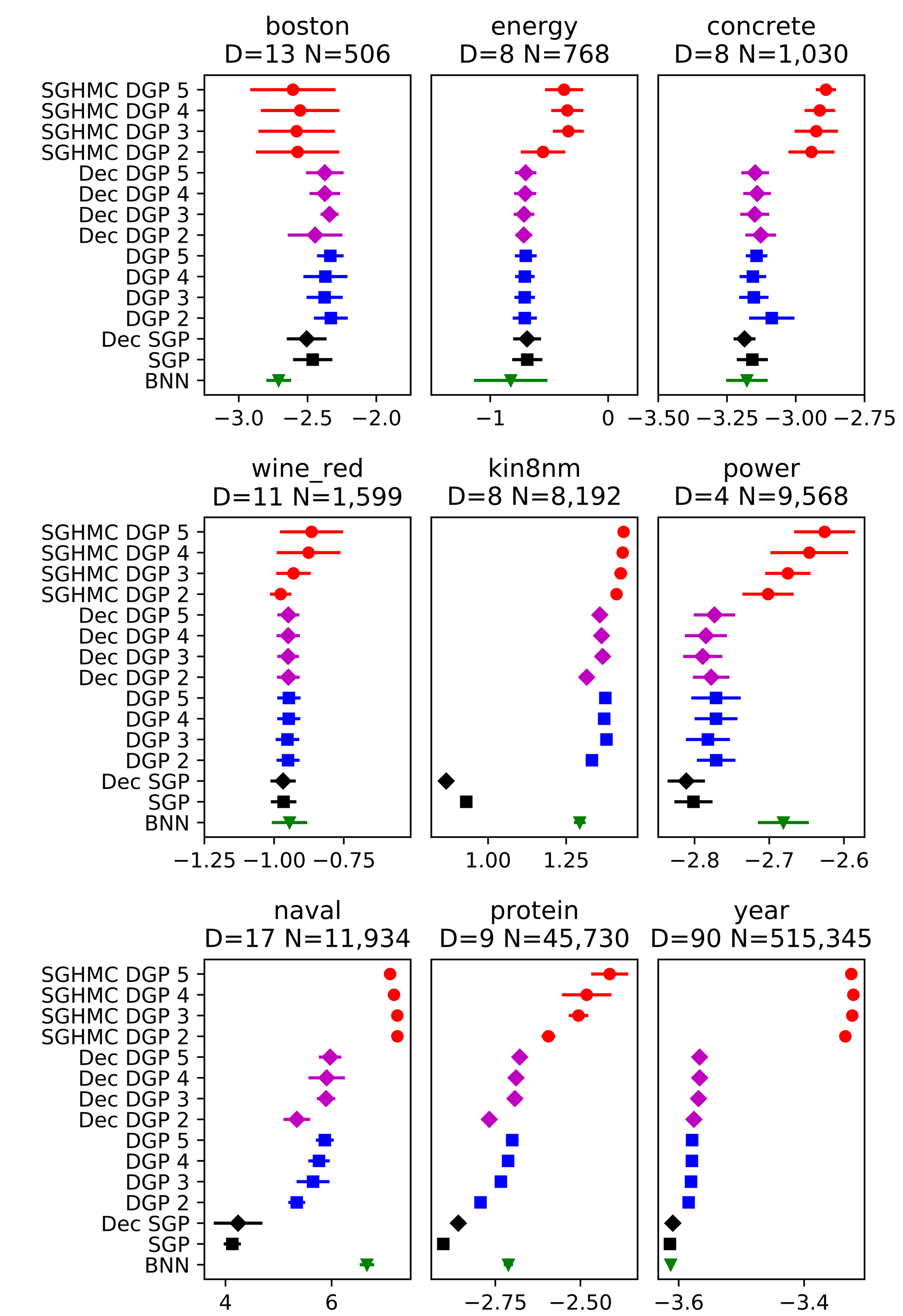


Figure: The test Mean Log-Likelihood on a variety of UCI datasets.

References

[1] Havasi, Marton, José Miguel Hernández-Lobato, and Juan José Murillo-Fuentes. "Inference in deep Gaussian processes using stochastic gradient Hamiltonian Monte Carlo." NeurIPS. 2018.

[2] Salimbeni, Hugh, and Marc Deisenroth. "Doubly stochastic variational inference for deep Gaussian processes." NeurIPS. 2017.

[3] Chen, Tianqi, Emily Fox, and Carlos Guestrin. "Stochastic gradient Hamiltonian Monte Carlo." ICML. 2014.

Contact: mh740@cam.ac.uk