Sparse Gaussian Process Hyperparameters: Optimize or Integrate?

Vidhi Lalchand

Wessel P. Bruinsma

David R. Burt

Carl E. Rasmussen

UNIVERSITY OF CAMBRIDGE

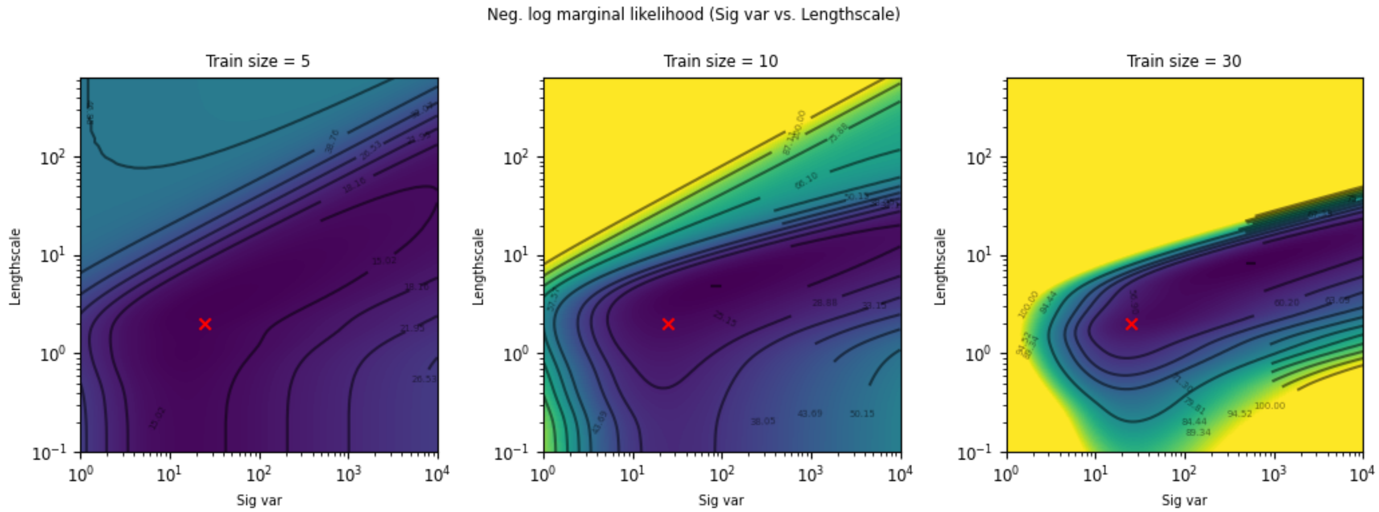
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- This work is about Bayesian hyperparameter inference in sparse Gaussian process regression.
- Traditional gradient based optimisation (ML-II) can be extremely sensitive to starting values.
- ML-II hyperparameter estimates are subject to high variability and underestimate prediction uncertainty.
- We propose a novel and computationally efficient scheme Fully Bayesian inference in sparse GPs.

$$y_n = f(x_n) + \epsilon_n, \ \epsilon_n \sim \mathcal{N}(0, \sigma^2), \ f \sim \mathcal{GP}(0, k_\theta)$$

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) = \log \int p(\boldsymbol{y}|f)p(f|\boldsymbol{\theta})df = c \underbrace{- \frac{1}{2}\boldsymbol{y}^T(K_\theta + \sigma^2 I)^{-1}\boldsymbol{y}}_{\text{data fit term}} - \underbrace{\frac{1}{2}|K_\theta + \sigma^2 I|}_{\text{complexity penalty}}$$



Mathematical set-up

Inputs $X = (\boldsymbol{x}_n)_{n=1}^N \subseteq \mathbb{R}^D$

Outputs $\boldsymbol{y} = (y_n)_{n=1}^N \subseteq \mathbb{R}$

Inducing variables $\boldsymbol{u} = \{f(\boldsymbol{z}_m)\}_{m=1}^M \subseteq \mathbb{R}$

Latent function prior $f \sim \mathcal{GP}(0, k_{\boldsymbol{\theta}})$

Factorised Gaussian likelihood $p(\boldsymbol{y}|f) = \prod_{n=1}^N \mathcal{N}(y_n|f_n, \sigma^2)$

Inducing locations $Z = \{\boldsymbol{z}_m\}_{m=1}^M, \boldsymbol{z}_m \in \mathbb{R}^D$

Canonical Inference for θ in sparse GPs:

1. Specify a variational approximation to the posterior over $(f, \boldsymbol{u})$
2. Lower bound the GP log-marginal likelihood $\log p(\boldsymbol{y}|\boldsymbol{\theta}) \geq \mathcal{L}_{\theta,Z}$
3. Use the closed-form ELBO to learn hyperparameters (θ) and inducing locations (Z)

Variational approximation to the posterior $p(f, \boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}) \approx q(f, \boldsymbol{u}|\boldsymbol{\theta}) = p(f|\boldsymbol{u}, \boldsymbol{\theta})q(\boldsymbol{u})$

Hyperparameter inference $\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta},Z} \mathcal{L}_{\boldsymbol{\theta},Z}.$ ← "Collapsed ELBO"

Titsias [2009] showed that in the case of a Gaussian likelihood the optimal variational distribution $q^*(u)$ is Gaussian and can be derived in closed-form.

[Ours] Doubly collapsed Inference for θ in sparse GPs:

1. Specify a variational approximation to the posterior over $(f, \boldsymbol{u}, \boldsymbol{\theta})$
2. Lower bound the GP log-marginal likelihood $\log p(\boldsymbol{y}) \geq \int q(\boldsymbol{\theta})\mathcal{L}_{\theta,Z} d\boldsymbol{\theta} - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}))$ ("Collapsed ELBO")
3. Crucially, we can write down the optimal $q^*(\boldsymbol{\theta})$ upto a normalising constant.

Variational approximation to the posterior $p(f, \boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{y}) \approx q(f, \boldsymbol{u}, \boldsymbol{\theta}) = p(f|\boldsymbol{u}, \boldsymbol{\theta})q(\boldsymbol{u}|\boldsymbol{\theta})q(\boldsymbol{\theta})$

Hyperparameter inference Sample $\boldsymbol{\theta}^* \sim q^*(\boldsymbol{\theta})$

Michalis Titsias. *Variational learning of inducing variables in sparse Gaussian processes.* In Artificial intelligence and statistics, pages 567–574. PMLR, 2009.

Algorithm

$$\log p(\boldsymbol{y}) \geq \int q(\boldsymbol{\theta})\mathcal{L}_{\boldsymbol{\theta},Z} d\boldsymbol{\theta} - \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}))$$

Overall, the core training algorithm alternates between two steps:

1. Sampling step for θ: $\theta_j \sim q^*(\theta) \propto \mathcal{L}(\boldsymbol{\theta}, Z_{opt}) + \log p(\boldsymbol{\theta})$, [Keep Z_{opt} fixed]
2. Optimisation step for Z: $Z_{opt} \longleftarrow \texttt{optim}(\hat{\mathcal{L}})$, where

$$\hat{\mathcal{L}} = \mathbb{E}_{q^*(\theta)}[\mathcal{L}(\theta, Z)] \approx \frac{1}{J}\sum_{j=1}^{J} \mathcal{L}(\theta_j, Z_{opt}),$$ [Keep θ fixed]

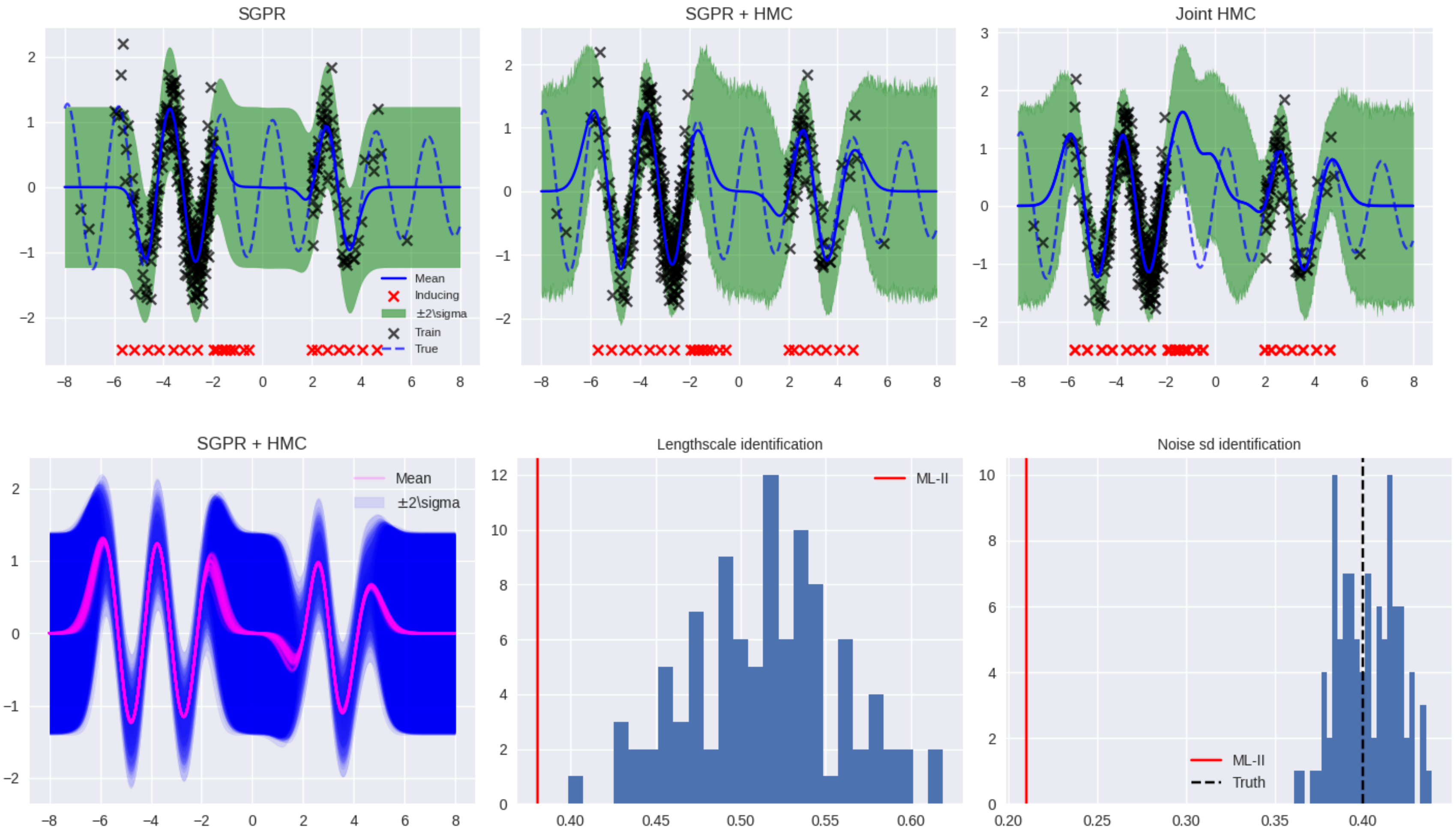
By sampling from $q^*(\boldsymbol{\theta})$, we side-step the need to sample from the joint $(\boldsymbol{u}, \boldsymbol{\theta})$-space yielding a significantly more efficient algorithm in the case of regression with a Gaussian likelihood.

Approach	Time/it.	Mem./it.	Param / Vars
Non-collapsed [Hensman et al, 2015]	m^3	m^2	$n_\theta + m$
Collapsed (ours)	nm^2	m^2	n_θ

n_θ is the number of hyperparameters and m is the number of inducing variables

James Hensman, Alexander G Matthews, Maurizio Filippone, and Zoubin Ghahramani. MCMC for variationally Sparse Gaussian processes. In Advances in Neural Information Porcessing Systems, pages 1648-1656, 2015.

1d Synthetic Experiment



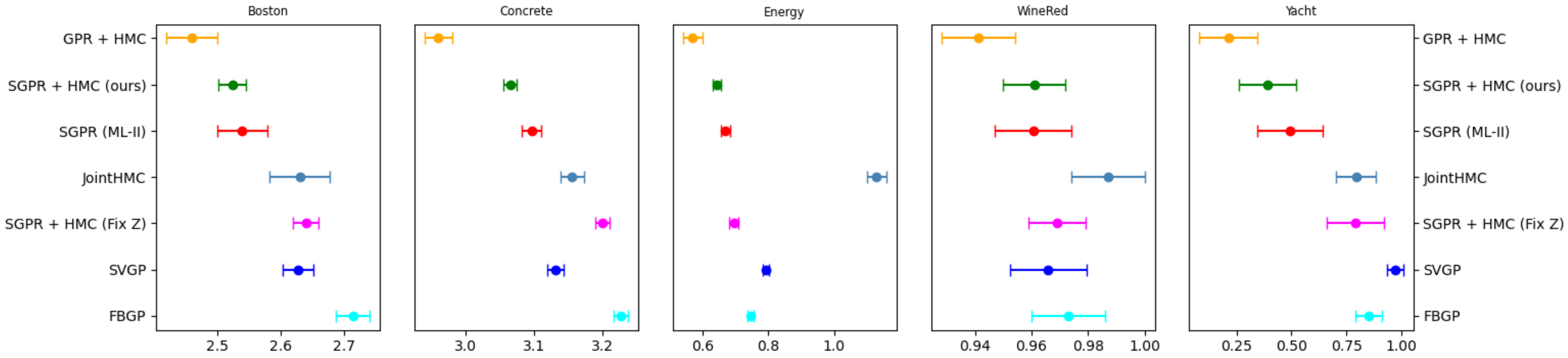
$$f(x) = \sin(3x) + 0.3\cos(\pi x)$$

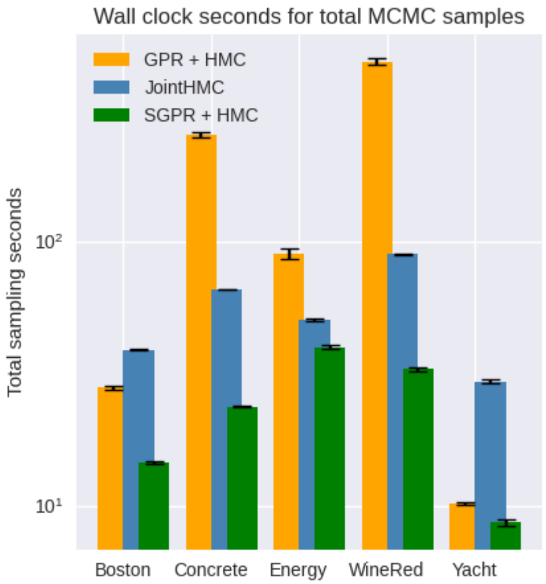
with the constraint $(x < -2)$ and $(x > 2)$.

Method	SGPR	SGPR + HMC	JointHMC
RMSE	0.580	**0.537**	0.682
NLPD	0.214	**0.065**	0.74

Sparse GP Benchmarks

Neg. log predictive density (mean ± se) on test data, 10 splits.





- Our method, SGPR + HMC (--) outperforms other fully Bayesian benchmarks like jointHMC (--) and FBGP (--) in terms of negative log predictive density on unseen data.
- It is significantly faster relative to jointHMC sand Exact GP inference with HMC (--).

Thank you!