Aalto University

Towards Improved Learning in Gaussian Processes: The Best of Both Worlds

Rui Li ST John Arno Solin

FCAI Finnish Center for Artificial Intelligence

TL;DR

Variational Inference (VI) and Expectation Propagation (EP) are two commonly used approximate inference methods in Gaussian Processes (GPs) which have complementary advantages. We design a hybrid training procedure to combine their respective advantages in capturing approximate posterior and approximate marginal likelihood, which can potentially result in improved hyperparameter learning.

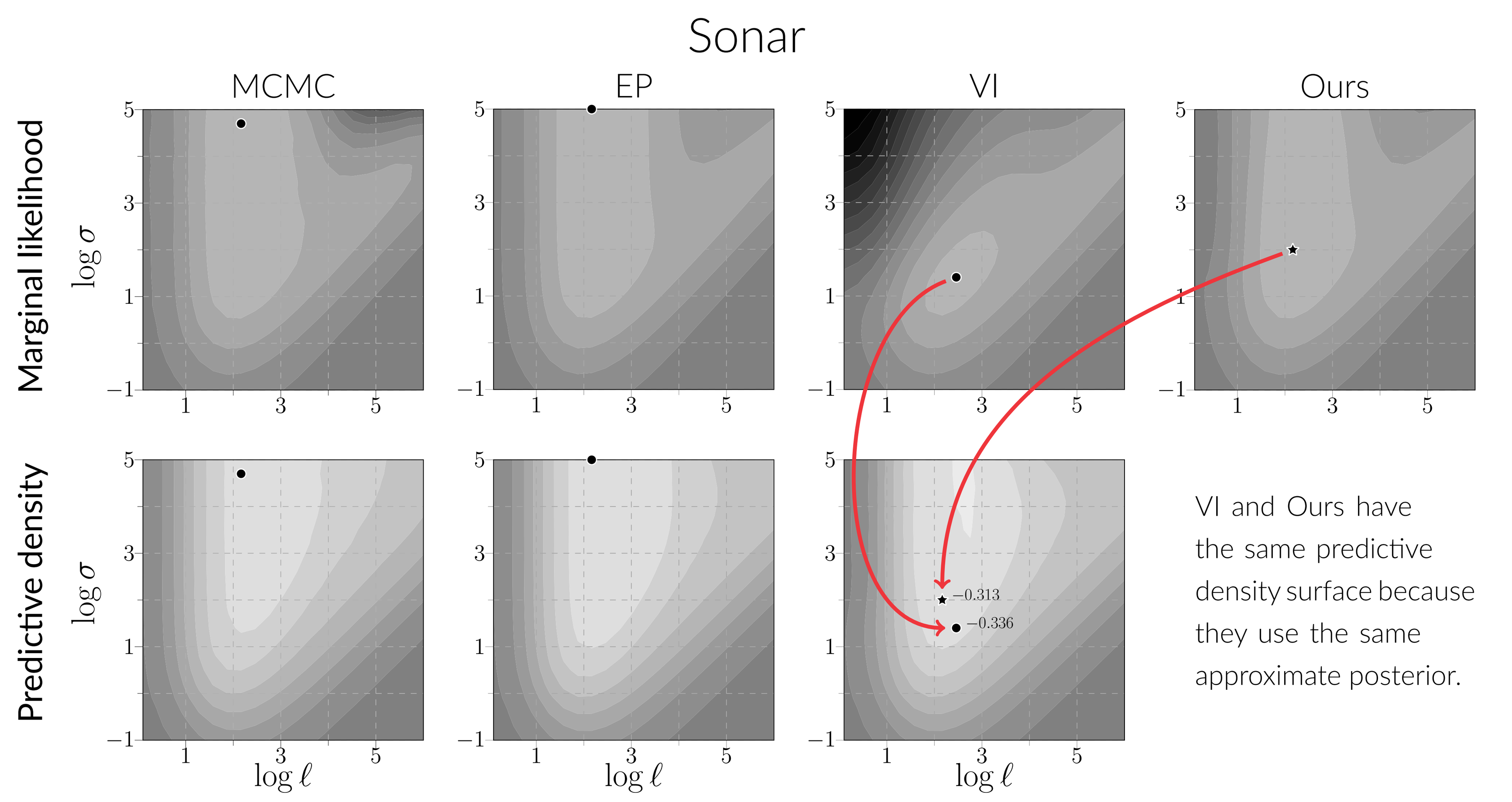

Figure 1. Colour scale is the same in all plots: -1.2 ▬ -0.2

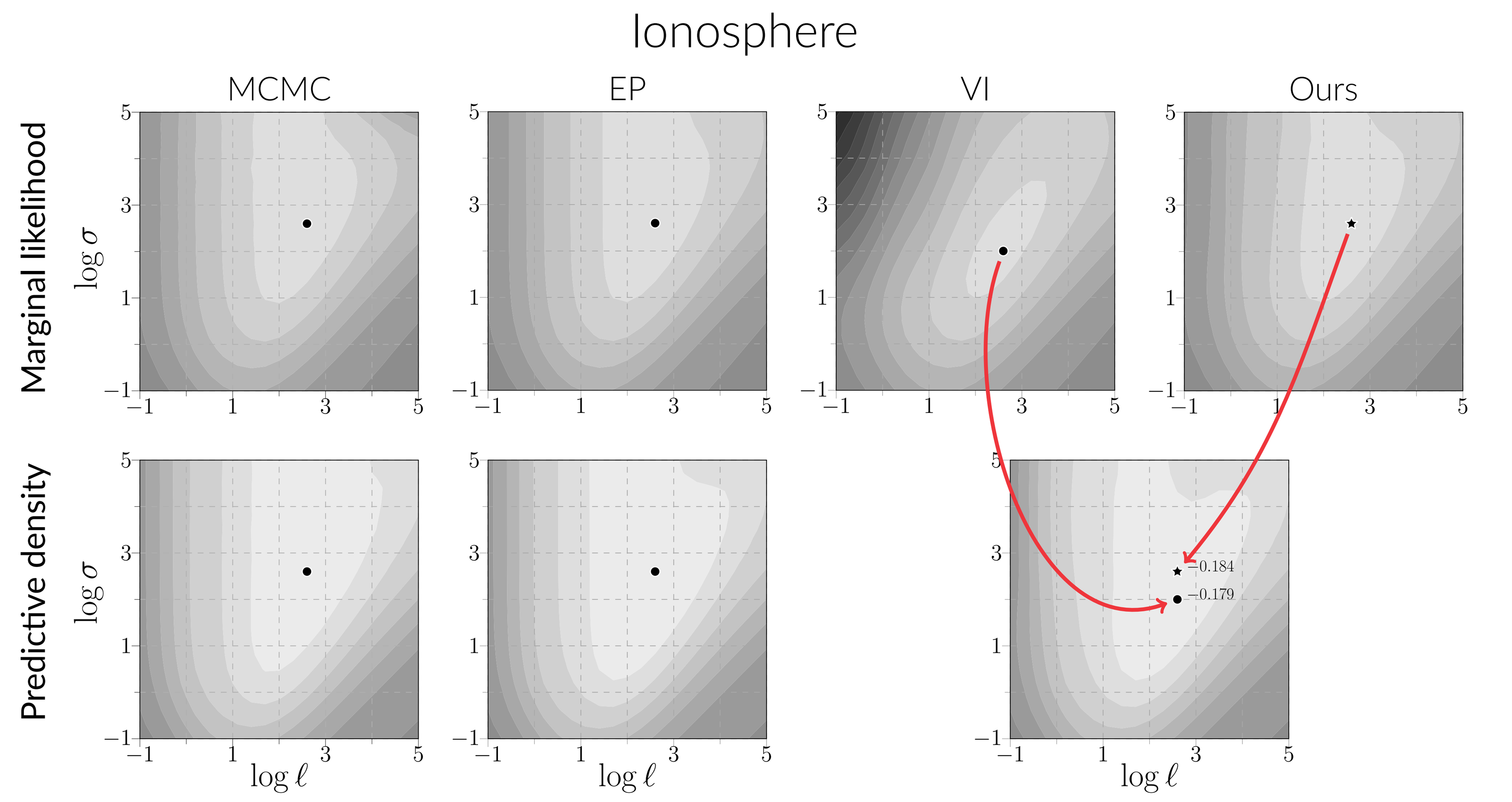

Figure 2. Colour scale is the same in all plots: -1.4 ▬ -0.1

Log marginal likelihood and predictive density surfaces for Sonar and Ionosphere (normalized by n). Black markers show optimal hyperparameter locations. EP and EP-like marginal likelihood estimation (Ours) match the MCMC baseline better than VI and result in better prediction.

Conjugate-computation Variational Inference Connects VI with EP

	Variational Inference	Expectation Propagation
Approximate Posterior	$q(\mathbf{f}; \boldsymbol{\lambda}, \boldsymbol{\theta}) \propto p(\mathbf{f}; \boldsymbol{\theta}) \prod_{i=1}^{n} \underbrace{\exp \langle \boldsymbol{\lambda}_i, \mathbf{T}(f_i) \rangle}_{t_i(f_i; \boldsymbol{\lambda}_i)}$	$q(\mathbf{f}; \boldsymbol{\zeta}, \boldsymbol{\theta}) = \frac{1}{Z} p(\mathbf{f}; \boldsymbol{\theta}) \prod_{i=1}^{n} t_i(f_i; \boldsymbol{\zeta}_i)$
Marginal Likelihood Estimation	$\mathcal{L}_{\mathsf{VI}}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = -\mathrm{D}_{\mathrm{KL}}[q(\mathbf{f}; \boldsymbol{\lambda}, \boldsymbol{\theta}) \parallel p(\mathbf{f}; \boldsymbol{\theta})] + \sum_{i=1}^{n} \mathbb{E}_{q(f_i; \boldsymbol{\lambda}_i, \boldsymbol{\theta})}[\log p(y_i \mid f_i; \boldsymbol{\theta})]$	$\mathcal{L}_{\mathsf{EP}}(\boldsymbol{\zeta}, \boldsymbol{\theta}) = \log \int p(\mathbf{f}; \boldsymbol{\theta}) \prod_{i=1}^{n} t_i(f_i; \boldsymbol{\zeta}_i) \, \mathrm{d}\mathbf{f}$

A new objective: $\mathcal{L}_{\mathsf{EP}}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \log \int p(\mathbf{f}; \boldsymbol{\theta}) \prod_{i=1}^{n} t_i(f_i; \boldsymbol{\lambda}_i) \, \mathrm{d}\mathbf{f}$

Improve Learning in GPs by Combining EP and VI

	Variational Expectation–Maximization Procedure	Hybrid Training Procedure	
E-step (Inference)	$\boldsymbol{\lambda}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\lambda}} \mathcal{L}_{\mathsf{VI}}(\boldsymbol{\lambda}, \boldsymbol{\theta}^{(t)})$	$\boldsymbol{\lambda}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\lambda}} \mathcal{L}_{\mathsf{VI}}(\boldsymbol{\lambda}, \boldsymbol{\theta}^{(t)})$	*numerically stable*
M-step (Learning)	$\boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{L}_{\mathsf{VI}}(\boldsymbol{\lambda}^{(t+1)}, \boldsymbol{\theta})$	$\boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{L}_{\mathsf{EP}}(\boldsymbol{\lambda}^{(t+1)}, \boldsymbol{\theta})$	*better estimation*

Experimental Evaluation: Binary Classification

- All GPs use Bernoulli likelihood and Matérn-$\frac{5}{2}$ kernel with hyperparameters $\boldsymbol{\theta} = (\ell, \sigma)$.
- We evaluate the log marginal likelihood on a regular 21×21 grid of values for $\log \boldsymbol{\theta}$ (see Figure 1 and 2).

Table 1. Test set accuracy and log predictive density on different data sets (mean ± standard deviation). Results that are statistically significantly different under a paired t-test ($p = 0.05$) are **bolded**.

	Accuracy		Log Predictive Density	
	VI	Ours	VI	Ours
Ionosphere	0.940 ± 0.016	0.946 ± 0.016	-0.179 ± 0.023	-0.176 ± 0.023
Sonar	0.836 ± 0.036	$\mathbf{0.860 \pm 0.034}$	-0.353 ± 0.013	$\mathbf{-0.340 \pm 0.015}$
Diabetes	0.783 ± 0.015	0.781 ± 0.013	-0.473 ± 0.030	-0.473 ± 0.030
USPS	0.974 ± 0.010	0.974 ± 0.010	-0.080 ± 0.011	$\mathbf{-0.077 \pm 0.011}$

References

[1] M. E. Khan and W. Lin, "Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models," in *AISTATS*, 2017.

[2] V. Adam, P. E. Chang, M. E. Khan, and A. Solin, "Dual parameterization of sparse variational Gaussian processes," in *NeurIPS*, 2021.

See paper PDF for details:
`arxiv.org/abs/2211.06260`

rui.li@aalto.fi