Learning from the dual parameterization

Approximate inference and learning in Gaussian process models

Ti John

Aalto University and Finnish Center for Artificial Intelligence

LIKE 23

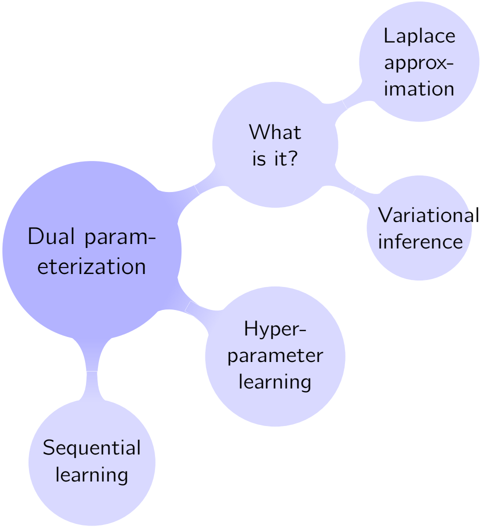
Bern, 28 June 2023

Joint work. . .

Rui Li

Paul E. Chang

Arno Solin


Dual param-
eterization
What
is it?
Laplace
approx-
imation
Variational
inference
Hyper-
parameter
learning
Sequential
learning

Outline

Gaussian process model

Laplace approximation

Variational inference

Hyperparameter learning

Sparse approximation

Sequential learning

Gaussian process model

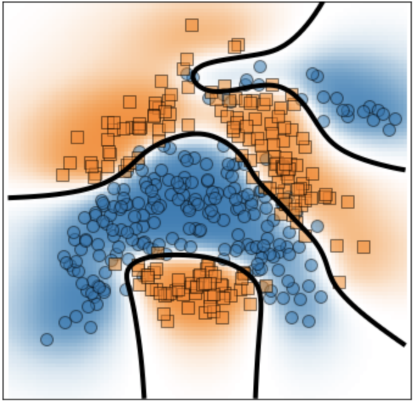
prior: $$p(f(\cdot)) = \mathcal{GP}(\mu(\cdot), \kappa(\cdot,\cdot))$$

likelihood: $$p(\mathbf{y} \,|\, \mathbf{f}) = \prod_{i=1}^{n} p(y_i \,|\, f_i), \quad \textit{e.g.}\ \text{Bernoulli for classification}$$

posterior: $$p(f(\cdot) \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \mathbf{f})\, p(f(\cdot)) \qquad \text{intractable}$$
$$\approx q(f(\cdot)) \qquad \text{approximate inference}$$

Approximate inference for non-conjugate likelihood models

- **MCMC (sampling) methods**
  (accurate but generally heavy)
- **Laplace approximation (LA)**
  (fast and simple)
- **Expectation propagation (EP)**
  (efficient but tricky)
- **Variational methods (VB/VI)**
  (popular but not problem-free)



GP classification with a Bernoulli likelihood

Gaussian approximate posterior

$$p(\mathbf{f} \mid \mathbf{y}) \approx q(\mathbf{f})$$

$$q(\mathbf{f}) = \mathrm{N}(\mathbf{f}; \mathbf{m}, \mathrm{S})$$

How to find **m** and S?

Laplace approximation

$$p(\mathbf{f} \mid \mathbf{y}) \approx q(\mathbf{f})$$

$$q(\mathbf{f}) = \mathrm{N}(\mathbf{f}; \underbrace{\mathbf{m}}_{\text{mode of } p(\mathbf{f} \mid \mathbf{y})}, \underbrace{\mathrm{S}}_{\text{determined by curvature at mode}})$$

How to find $\mathbf{m}$ and S?

"Point estimate++"

Laplace approximation, mean parameter **m**

posterior:

$$p(\mathbf{f} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f})$$

Laplace objective:

$$\mathcal{L}_{\mathrm{Lap}} = \log p(\mathbf{y} \mid \mathbf{f}) + \log p(\mathbf{f})$$

mode of posterior:

$$\mathbf{m} = \mathbf{f}^* = \arg\max_{\mathrm{f}} \mathcal{L}_{\mathrm{Lap}}(\mathbf{f})$$

(log-concave likelihoods: convex optimization, unique global maximum)

We can find a different (**dual**) parameterization!

Stationary point of $\mathcal{L}_{\mathrm{Lap}} = \log p(\mathbf{y} \mid \mathbf{f}) + \log p(\mathbf{f})$

At the optimum: $\log p(\mathbf{f}) = -\frac{1}{2}\mathbf{f}^\top \mathrm{K}^{-1}\mathbf{f} + \text{const.}$

$$
\begin{aligned}
0 &= \nabla_{\mathrm{f}} \mathcal{L}_{\mathrm{Lap}} \Big|_{\mathrm{f}=\mathrm{f}^*} \\
&= \underbrace{\nabla_{\mathrm{f}} \log p(\mathbf{y} \mid \mathbf{f})}_{\text{likelihood term}} \Big|_{\mathrm{f}=\mathrm{f}^*} + \underbrace{\nabla_{\mathrm{f}} \log p(\mathbf{f})}_{\text{prior term}} \Big|_{\mathrm{f}=\mathrm{f}^*} \\
&= \underbrace{\nabla_{\mathrm{f}} \log p(\mathbf{y} \mid \mathbf{f}) \Big|_{\mathrm{f}=\mathrm{f}^*}}_{=:\boldsymbol{\alpha}(\mathrm{f}^*)} - \mathrm{K}^{-1}\mathbf{f} \Big|_{\mathrm{f}=\mathrm{f}^*}
\end{aligned}
$$

$$
0 = \boldsymbol{\alpha}^* - \mathrm{K}^{-1}\mathbf{f}^* \qquad \Leftrightarrow \qquad \mathbf{f}^* = \mathrm{K}\boldsymbol{\alpha}^*
$$

$\Rightarrow$ mode can equivalently be parameterized through the derivatives of the likelihood

$\Rightarrow$ dual parameterization

What about the covariance S?

Laplace approximation: 2nd-order Taylor approximation to log posterior.

$$\text{precision} \quad \mathrm{S}^{-1} = -\nabla^2_{\mathrm{f}} \mathcal{L}_{\mathrm{Lap}}\big|_{\mathrm{f}=\mathrm{f}^*}$$

Hessian of negative log posterior, $-\mathcal{L}_{\mathrm{Lap}} = -\log p(\mathbf{y} \mid \mathbf{f}) - \log p(\mathbf{f})$:

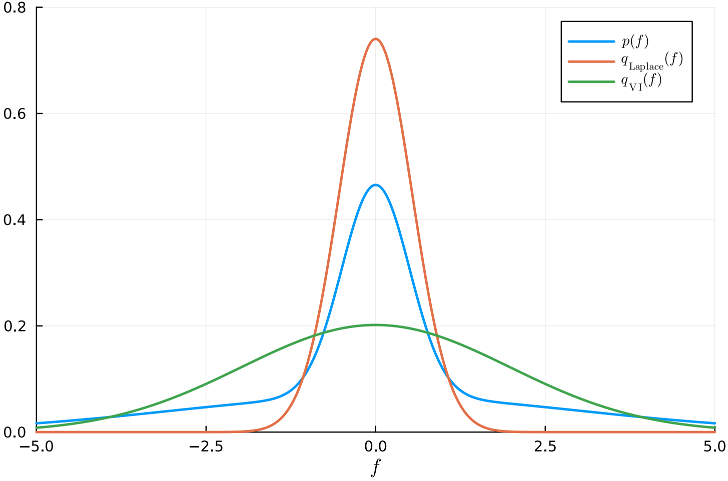
$$-\nabla^2_{\mathrm{f}} \mathcal{L}_{\mathrm{Lap}}\big|_{\mathrm{f}=\mathrm{f}^*} = \underbrace{-\nabla^2_{\mathrm{f}} \log p(\mathbf{y} \mid \mathbf{f})\big|_{\mathrm{f}=\mathrm{f}^*}}_{=:\mathrm{W}} + \mathrm{K}^{-1}$$

For factorizing likelihood, $p(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{n} p(y_i \mid f_i)$:

$$\mathrm{W} = -\nabla^2_{\mathrm{f}} \log p(\mathbf{y} \mid \mathbf{f})\big|_{\mathrm{f}=\mathrm{f}^*} = \mathrm{diag}(\boldsymbol{\beta}), \qquad \beta_i = -\frac{\partial^2}{\partial f_i^2} \log p(y_i \mid f_i)$$

$$\Rightarrow \mathrm{S}_{\mathrm{Lap}} = (\mathrm{W} + \mathrm{K}^{-1})^{-1}$$

Laplace approximation is local



Instead of point estimate (and post-hoc uncertainty), we may prefer optimizing over a whole posterior distribution directly ⇒ variational inference (VI)

Dual parameterization of VI

$$q(\mathbf{f}) = \mathrm{N}(\mathbf{f}; \mathbf{m}, \mathrm{S})$$

sufficient statistics (f_i, f_i^2)

$$q^*(\mathbf{f}) \propto p(\mathbf{f}) \prod_{i=1}^{n} \underbrace{\exp(\langle \lambda_i^*, T(f_i) \rangle)}_{\text{1D sites}} \qquad \text{where } \lambda_i^* = \underbrace{\nabla_{\mu_i} \mathbb{E}_{q^*(f_i)}[\log p(y_i \mid f_i)]}_{\text{nat.grad.}}$$

(α_i, β_i)

From Laplace to VI

$$\mathcal{L}_{\text{Lap}}[\mathbf{f}] = \underbrace{\log p(\mathbf{y} \mid \mathbf{f})}_{\text{data fit}} + \underbrace{\log p(\mathbf{f})}_{\text{regularizer}}$$

$$\mathcal{L}_{\text{VI}}[q(\mathbf{f})] = \mathbb{E}_{q(\mathrm{f})} \log p(\mathbf{y} \mid \mathbf{f}) + \mathbb{E}_{q(\mathrm{f})} \log p(\mathbf{f}) + \mathbb{H}[q(\mathbf{f})]$$

$$= \mathbb{E}_{q(\mathrm{f})} \log p(\mathbf{y} \mid \mathbf{f}) - \underbrace{\mathbb{E}_{q(\mathrm{f})} \log \frac{q(\mathbf{f})}{p(\mathbf{f})}}_{\mathrm{KL}\left[q(\mathrm{f}) \| p(\mathrm{f})\right]}$$

entropy
$\Rightarrow$ valid objective
($\rightarrow$ Generalized VI)

$$\mathcal{L}_{\text{ELBO}}[q(\mathbf{f})] = \mathbb{E}_{q(\mathrm{f})} \log p(\mathbf{y} \mid \mathbf{f}) - \mathrm{KL}\left[q(\mathbf{f}) \| p(\mathbf{f})\right]$$

Stationary point

$$q^*(\mathbf{f}) = \arg\max_q \mathcal{L}_{\mathrm{ELBO}}[q] \qquad \text{for } q(\mathbf{f}) = \mathrm{N}(\mathbf{m}, \mathsf{S})$$

Now two stationary point equations:

$$0 = \nabla_{\mathsf{m}} \mathcal{L}_{\mathrm{ELBO}}$$
$$0 = \nabla_{\mathsf{S}} \mathcal{L}_{\mathrm{ELBO}}$$

Equation for mean $\mathbf{m}$

$$\nabla_{\mathrm{m}} \mathcal{L}_{\mathrm{ELBO}}[q(\mathbf{f})] = \nabla_{\mathrm{m}} \mathbb{E}_{q(\mathrm{f})}[\log p(\mathbf{y} \mid \mathbf{f})] - \nabla_{\mathrm{m}} \mathrm{KL}\left[q(\mathbf{f}) \| p(\mathbf{f})\right]$$

$$\mathrm{KL}\left[q(\mathbf{f}) \| p(\mathbf{f})\right] = \mathrm{KL}\left[\mathrm{N}(\mathbf{m}, \mathrm{S}) \| \mathrm{N}(\mathbf{0}, \mathrm{K})\right] = \frac{1}{2}\left(\mathrm{Tr}(\mathrm{K}^{-1}\mathrm{S}) - n + \mathbf{m}^{\mathsf{T}}\mathrm{K}^{-1}\mathbf{m} + \log \frac{\det \mathrm{K}}{\det \mathrm{S}}\right)$$

$$\nabla_{\mathrm{m}} \mathrm{KL}\left[q(\mathbf{f}) \| p(\mathbf{f})\right] = \mathrm{K}^{-1}\mathbf{m}$$

$$\nabla_{\mathrm{m}} \mathbb{E}_{q(\mathrm{f})}[\log p(\mathbf{y} \mid \mathbf{f})] \underset{\text{Bonnet's theorem}}{=} \mathbb{E}_{q(\mathrm{f})}[\nabla_{\mathrm{f}} \log p(\mathbf{y} \mid \mathbf{f})] =: \boldsymbol{\alpha}$$

At optimum:

$$0 = \boldsymbol{\alpha}^* - \mathrm{K}^{-1}\mathbf{m}^* \qquad \Leftrightarrow \qquad \mathbf{m}^* = \mathrm{K}\boldsymbol{\alpha}^*$$

Equation for covariance S

more complicated. . .

Reparameterizations

$$q(\mathbf{f}) = \mathrm{N}(\mathbf{m}, \mathsf{S})$$

- mean–covariance: $\xi = (\mathbf{m}, \mathsf{S})$ (and whitened reparameterization)
- natural parameters: $\eta = (\mathsf{S}^{-1}\mathbf{m}, -\frac{1}{2}\mathsf{S}^{-1})$
- expectation parameters: $\mu = (\mathbf{m}, \mathbf{m}\mathbf{m}^\top + \mathsf{S})$

Lagrangian dual

$$\mathcal{L}_{\text{ELBO}}[q(\mathbf{f})] = \mathbb{E}_{q(\mathbf{f})} \log p(\mathbf{y} \mid \mathbf{f}) - \mathrm{KL}\left[q(\mathbf{f}) \| p(\mathbf{f})\right]$$

Introducing local $\tilde{\mu}$ and moment-matching constraint:

$$\mathcal{L}_{\text{Lagrange}}(\tilde{\mu}, \mu, \lambda) = \sum_{i=1}^{n} \mathbb{E}_{\tilde{q}_i}(f_i; \tilde{\mu}_i)[\log p(y_i \mid f_i)] - \sum_{i=1}^{n} \langle \lambda_i, \tilde{\mu}_i - \mu_i \rangle - \mathrm{KL}\left[q(\mathbf{f}; \mu) \| p(\mathbf{f})\right]$$

Stationary point: $q^*(\mathbf{f})$ has natural parameters $\eta_q^* = \eta_p + \lambda^*$

$$\begin{aligned}
0 &= \nabla_\lambda \mathcal{L}_{\text{Lagrange}} & \Rightarrow \quad \mu^* &= \tilde{\mu}^* \\
0 &= \nabla_{\tilde{\mu}} \mathcal{L}_{\text{Lagrange}} & \Rightarrow \quad \lambda_i^* &= \nabla_{\mu_i} \mathbb{E}_{q^*(f_i)}[\log p(y_i \mid f_i)] \\
0 &= \nabla_\mu \mathcal{L}_{\text{Lagrange}} & \Rightarrow \quad \lambda^* &= \underbrace{\nabla_\mu \mathrm{KL}\left[q(\mathbf{f}; \mu^*) \| p(\mathbf{f})\right]}_{\eta_q - \eta_p}
\end{aligned}$$

Optimal $q^*(\mathbf{f})$

$$\text{Optimal } q^*(\mathbf{f}) \text{ has natural parameters } \eta_q^* = \eta_p + \lambda^*$$
$$\text{Prior } p(\mathbf{f}) \text{ has natural parameters } \eta_p = (0, -\frac{1}{2}\mathrm{K}^{-1})$$

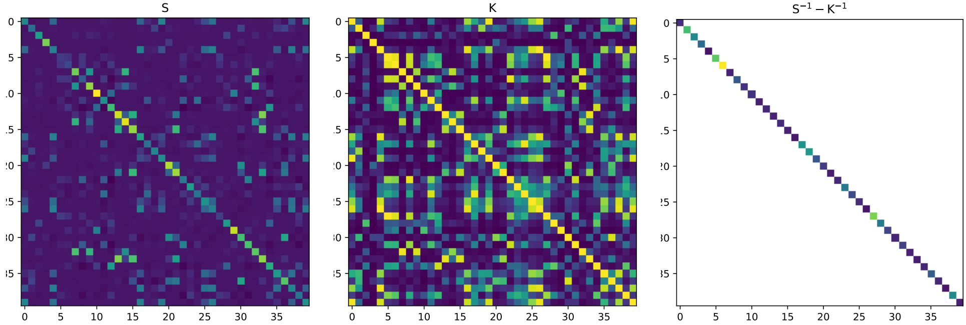
In mean–covariance parameterization:

$$\mathbf{m}^* = \mathrm{K}\boldsymbol{\alpha}^* \qquad \boldsymbol{\alpha}^* = \mathbb{E}_{q^*(\mathrm{f})}[\nabla_{\mathrm{f}} \log p(\mathbf{y} \mid \mathbf{f})]$$
$$(\mathrm{S}^*)^{-1} = \mathrm{K}^{-1} + \boldsymbol{\beta}^* \qquad \boldsymbol{\beta}^* = \mathbb{E}_{q^*(\mathrm{f})}[-\nabla_{\mathrm{f}}^2 \log p(\mathbf{y} \mid \mathbf{f})]$$

$\Rightarrow$ optimal Gaussian approximate posterior for factorizing likelihoods:

$$q^*(\mathbf{f}) = \frac{1}{Z} p(\mathbf{f}) \prod_{i=1}^{n} t_i(f_i)$$

Optimal posterior decomposition



How to find $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}_i^*$?

Natural gradient updates:

$$\lambda^{k+1} = (1-\rho)\lambda^k + \underbrace{\rho}_{\text{learning rate}}\nabla_\mu \mathbb{E}_{q(\mathrm{f})}[\log p(\mathbf{y} \mid \mathbf{f})]$$

- cheap: same cost as standard gradient descent, no dense Hessian required!

Learning hyperparameters θ

$$p(\theta \,|\, \mathcal{D}) \propto \underbrace{p(\mathcal{D} \,|\, \theta)}_{\int p(\mathcal{D} \,|\, \mathrm{f})\, p(\mathrm{f} \,|\, \theta)\, \mathrm{df}}\ p(\theta)$$

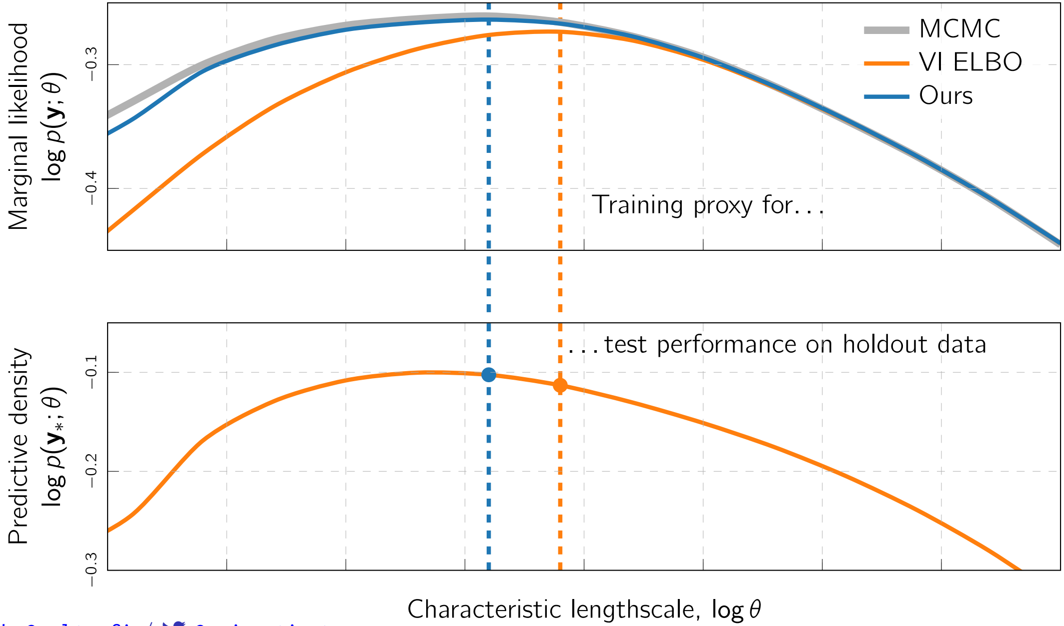
- $p(\theta \,|\, \mathcal{D})$: e.g. MCMC
- point estimate θ^*
  - maximum likelihood: $\theta^* = \arg\max_\theta p(\mathcal{D} \,|\, \theta)$

Marginal likelihood

$$\log p(\mathcal{D} \mid \theta) = \log \int p(\mathcal{D} \mid \mathbf{f})\, p(\mathbf{f} \mid \theta)\, \mathrm{d}\mathbf{f}$$

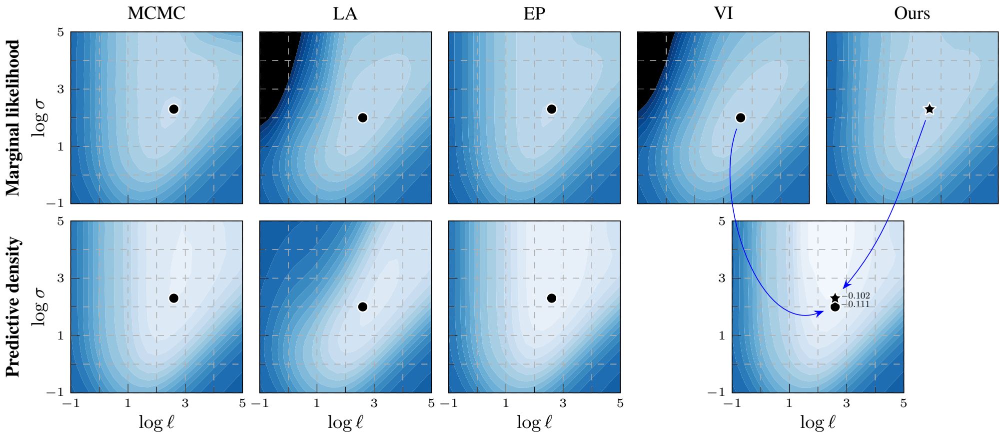
$$\begin{aligned}
\text{Laplace:} \qquad & \log p(\mathbf{y} \mid \theta) = \log \int \exp(\mathcal{L}_{\mathrm{Lap}}(\mathbf{f}))\, \mathrm{d}\mathbf{f} \\
& \qquad \approx \mathcal{L}_{\mathrm{Lap}}(\mathbf{f}^*) - \frac{1}{2} \log \det \mathrm{S}_{\mathrm{Lap}} + \text{const.} \\
\text{VI:} \qquad & \log p(\mathbf{y} \mid \theta) \geq \mathcal{L}_{\mathrm{ELBO}}[q^*] \\
\text{EP:} \qquad & \log p(\mathbf{y} \mid \theta) \approx \mathcal{L}_{\mathrm{EP}}[q^*] \\
& \qquad = \log \int p(\mathbf{f}) \prod_{i=1}^{n} t(f_i)\, \mathrm{d}\mathbf{f}
\end{aligned}$$

Same form as dual parameterization!



MCMC
VI ELBO
Ours
Training proxy for...
...test performance on holdout data
Marginal likelihood
$\log p(\mathbf{y};\theta)$
−0.3
−0.4
Predictive density
$\log p(\mathbf{y}_*;\theta)$
−0.1
−0.2
−0.3
Characteristic lengthscale, $\log\theta$

Marginal likelihood estimation



Variational Expectation–Maximization

E-step (inference): $\boldsymbol{\lambda}^{(k+1)} \leftarrow \arg\max_{\boldsymbol{\lambda}} \mathcal{L}_{\mathrm{E}}(\boldsymbol{\lambda}, \boldsymbol{\theta}^{(k)})$

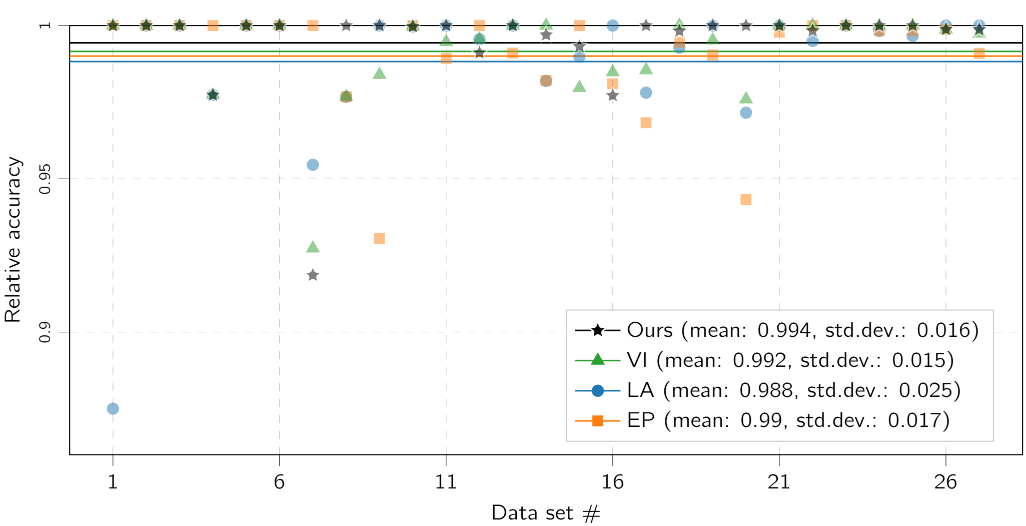
M-step (learning): $\boldsymbol{\theta}^{(k+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{M}}(\boldsymbol{\lambda}^{(k+1)}, \boldsymbol{\theta})$

$\mathcal{L}_{\mathrm{E}} \equiv \mathcal{L}_{\text{ELBO}}$

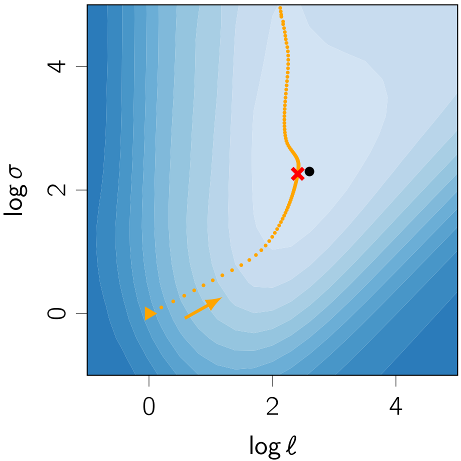
$\mathcal{L}_{\mathrm{M}} \equiv \mathcal{L}_{\text{EP}}$

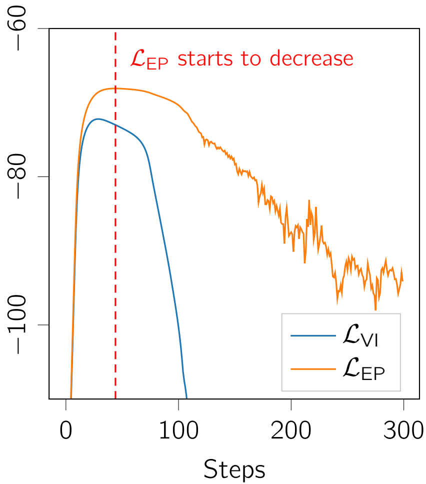
	$(n,\ d)$	LA	EP	VI	Ours	MCMC
trains	(10, 30)	−0.702±0.025	−0.698±0.033	−0.702±0.037	−0.691±0.046	−0.692±0.025
balloons	(16, 5)	−0.660±0.125	−0.650±0.128	−0.649±0.185	−0.607±0.227	−0.684±0.076
fertility	(100, 10)	−0.388±0.122	−0.384±0.149	−0.393±0.136	−0.397±0.139	−0.382±0.126
pittsburg-bridges-T-OR-D	(102, 8)	−0.299±0.081	−0.321±0.108	−0.290±0.110	−0.293±0.116	−0.306±0.115
acute-nephritis	(120, 7)	−0.203±0.012	−0.046±0.007	−0.007±0.002	−0.005±0.002	−0.005±0.002
acute-inflammation	(120, 7)	−0.184±0.018	−0.052±0.007	−0.007±0.002	−0.007±0.002	−0.007±0.003
echocardiogram	(131, 11)	−0.424±0.093	−0.418±0.095	−0.425±0.110	−0.428±0.112	−0.437±0.127
hepatitis	(155, 20)	−0.370±0.071	−0.372±0.072	−0.364±0.090	−0.367±0.094	−0.369±0.091
parkinsons	(195, 23)	−0.260±0.031	−0.295±0.056	−0.160±0.050	−0.141±0.046	−0.145±0.044
breast-cancer-wisc-prog	(198, 34)	−0.458±0.075	−0.473±0.091	−0.457±0.085	−0.460±0.088	−0.464±0.085
spect	(265, 23)	−0.593±0.049	−0.590±0.055	−0.594±0.054	−0.595±0.054	−0.596±0.051
statlog-heart	(270, 14)	−0.395±0.064	−0.389±0.061	−0.396±0.071	−0.397±0.071	−0.397±0.070
haberman-survival	(306, 4)	−0.530±0.053	−0.532±0.059	−0.531±0.055	−0.531±0.055	−0.520±0.063
ionosphere	(351, 34)	−0.224±0.042	−0.230±0.042	−0.170±0.048	−0.170±0.055	−0.179±0.058
horse-colic	(368, 26)	−0.463±0.059	−0.452±0.057	−0.467±0.072	−0.473±0.082	−0.469±0.079
congressional-voting	(435, 17)	−0.640±0.028	−0.639±0.030	−0.641±0.030	−0.642±0.029	−0.644±0.027
cylinder-bands	(512, 36)	−0.488±0.038	−0.500±0.041	−0.465±0.049	−0.451±0.052	−0.451±0.049
breast-cancer-wisc-diag	(569, 31)	−0.085±0.026	−0.140±0.020	−0.077±0.044	−0.075±0.045	−0.076±0.043
ilpd-indian-liver	(583, 10)	−0.513±0.040	−0.520±0.041	−0.512±0.043	−0.512±0.043	−0.512±0.042
monks-2	(601, 7)	−0.491±0.025	−0.512±0.028	−0.464±0.031	−0.442±0.033	−0.437±0.032
statlog-australian-credit	(690, 15)	−0.630±0.026	−0.639±0.036	−0.630±0.026	−0.630±0.026	−0.630±0.025
credit-approval	(690, 16)	−0.342±0.047	−0.342±0.050	−0.341±0.052	−0.342±0.052	−0.341±0.052
breast-cancer-wisc	(699, 10)	−0.094±0.025	−0.093±0.023	−0.093±0.029	−0.093±0.029	−0.093±0.029
blood	(748, 5)	−0.478±0.039	−0.479±0.040	−0.478±0.039	−0.478±0.039	−0.478±0.039
pima	(768, 9)	−0.474±0.033	−0.476±0.038	−0.474±0.035	−0.474±0.035	−0.474±0.035
mammographic	(961, 6)	−0.407±0.038	−0.407±0.040	−0.408±0.040	−0.408±0.040	−0.408±0.040
statlog-german-credit	(1000, 25)	−0.491±0.030	−0.491±0.032	−0.492±0.032	−0.492±0.032	−0.492±0.032
Bold Count		14	13	13	16	/

Relative accuracy (compared to best method on each data set)



Optimization issues. . .





What about big data?

Sparse approximation using $\mathbf{u} = f(\mathrm{Z})$:

$$q_{\mathrm{u}}(f(\cdot); \boldsymbol{\xi}_{\mathrm{u}}, \boldsymbol{\theta}) = \int p(f(\cdot) \mid \mathbf{u}; \boldsymbol{\theta})\, q(\mathbf{u}; \boldsymbol{\xi}_{\mathrm{u}})\, \mathrm{d}\mathbf{u},$$

Dual parameters:

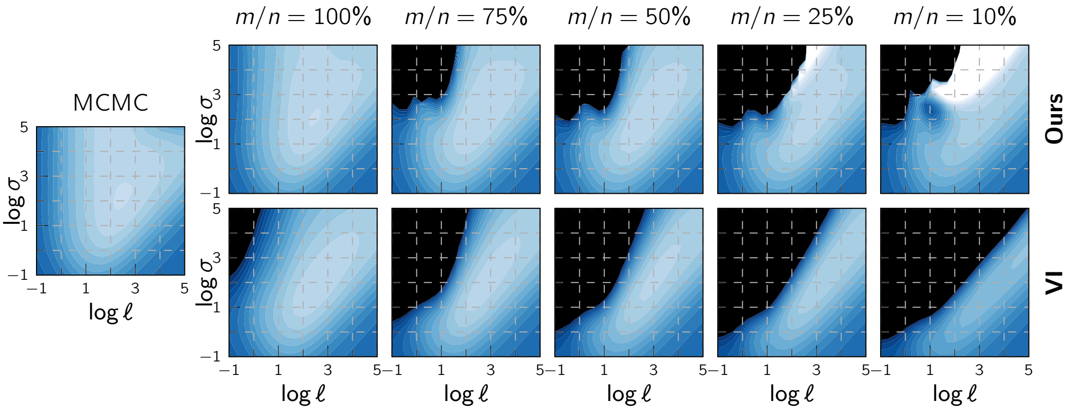
$$\hat{\alpha}_i = \mathbb{E}_{q_{\mathrm{u}}(f_i)}[\nabla_{f_i} \log p(y_i \mid f_i)] \qquad \hat{\beta}_i = \mathbb{E}_{q_{\mathrm{u}}(f_i)}[-\nabla^2_{f_i} \log p(y_i \mid f_i)]$$

Projection onto sparse inducing points:

$$\boldsymbol{\alpha}_{\mathrm{u}} = \sum_{i=1}^{n} \mathbf{k}_{\mathrm{z}_i}\, \hat{\alpha}_i \qquad \mathbf{B}_{\mathrm{u}} = \sum_{i=1}^{n} \mathbf{k}_{\mathrm{z}_i}\, \hat{\beta}_i\, \mathbf{k}_{\mathrm{z}_i}^{\top}$$

vector evaluated from kernel $\kappa(\mathbf{x}_i, \mathbf{z}_j)$

Sparse marginal likelihood approximations



Additive structure of dual parameterization in sequential learning

$$\hat{\alpha}_i = \mathbb{E}_{q_u(f_i)}[\nabla_{f_i} \log p(y_i \mid f_i)] \qquad \hat{\beta}_i = \mathbb{E}_{q_u(f_i)}[-\nabla^2_{f_i} \log p(y_i \mid f_i)]$$

$$\boldsymbol{\alpha}_u = \sum_{i=1}^{n} \mathbf{k}_{z_i} \hat{\alpha}_i \qquad \mathbf{B}_u = \sum_{i=1}^{n} \mathbf{k}_{z_i} \hat{\beta}_i \mathbf{k}_{z_i}^\top$$

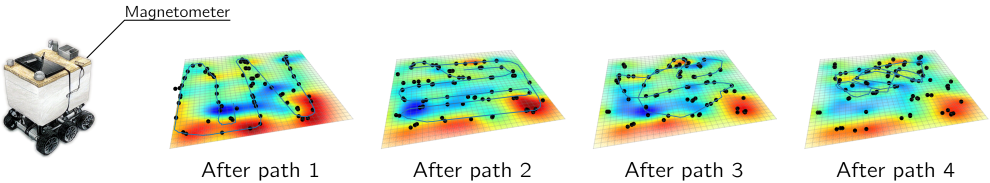
$$\mathcal{L}_{\text{batch}}(\mathbf{m}, \mathrm{S} \mid \mathcal{D}) = \sum_{i \in \mathcal{D}} \mathbb{E}_{q_u(f_i)}[\log p(y_i \mid f_i)] - \mathrm{KL}\left[q_u(\mathbf{u}) \| p_{\boldsymbol{\theta}}(\mathbf{u})\right]$$

$$\mathcal{L}_{\text{batch}}(\mathbf{m}, \mathrm{S} \mid \cancel{\mathcal{D}_{\mathrm{old}}} \cup \mathcal{D}_{\mathrm{new}}) = \sum_{i \in \cancel{\mathcal{D}_{\mathrm{old}}} \cup \mathcal{D}_{\mathrm{new}}} \mathbb{E}_{q_u(f_i)}[\log p(y_i \mid f_i)] - \mathrm{KL}\left[q_u(\mathbf{u}) \| p_{\boldsymbol{\theta}}(\mathbf{u})\right]$$

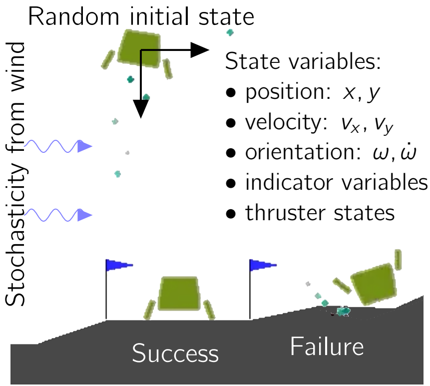
$$\boldsymbol{\alpha}_u = \boldsymbol{\alpha}_u^{\text{old}} + \boldsymbol{\alpha}_u^{\text{new}} \qquad \text{and} \qquad \mathbf{B}_u = \mathbf{B}_u^{\text{old}} + \mathbf{B}_u^{\text{new}}$$

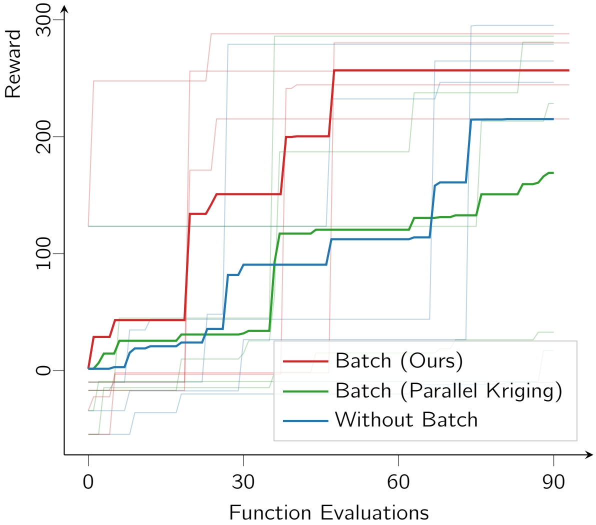
- natural gradient descent on new batch to find $\boldsymbol{\alpha}_u^{\text{new}}$ and $\mathbf{B}_u^{\text{new}}$

Continual learning



Bayesian optimization with fantasizing





Learning from the dual parameterization

- Dual parameters: derivatives of log likelihood ⇔ sensitivities w.r.t. data points
- Cheap natural gradient updates
- EP-like objective for hyperparameter learning
- Good parameterization for sequential learning

Find out more:

- *Improving Hyperparameter Learning under Approximate Inference in Gaussian Process Models.* Li, John, & Solin; ICML 2023. (arXiv:2306.04201)
- *Memory-Based Dual Gaussian Processes for Sequential Learning.* Chang, Verma, John, Solin, & Khan; ICML 2023. (arXiv:2306.03566)
- *Dual parameterization for dummies* (in preparation, coming to an arXiv near you)