A APPENDIX

A.1 Proof of theorem 2

Theorem 2 states:

Theorem. *The **complete conditional distributions** of the augmented model presented in Section 3.1 are given by*

$$\begin{aligned} p(\omega_i|f_i, y_i) =& \pi_\varphi\left(\omega_i|\|h(f_i, y_i)\|_2\right), \\ p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{\omega}) =& \mathcal{N}\left(\boldsymbol{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \end{aligned}$$

where $\boldsymbol{\Sigma} = \left(\operatorname{diag}\left(2\boldsymbol{\omega} \circ \gamma(\boldsymbol{y})\right) + K^{-1}\right)^{-1}$ *and* $\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(g(\boldsymbol{y}) + \boldsymbol{\omega} \circ \beta(\boldsymbol{y}) + K^{-1}\boldsymbol{\mu}_0\right)$, $\circ$ *denotes the Hadamard product and the function* $h(\cdot)$ *is given by the form of likelihood (see Eq.5).*

Proof: For the full conditional on $\boldsymbol{f}$:

$$\begin{aligned} p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{\omega}) \propto& p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{\omega})p(\boldsymbol{f}) \\ \propto& \exp\left[g(\boldsymbol{y})^\top \boldsymbol{f} + (\beta(\boldsymbol{y}) \circ \boldsymbol{\omega})^\top \boldsymbol{f} - \boldsymbol{f}^\top \operatorname{diag}(\gamma(\boldsymbol{y}) \circ \boldsymbol{\omega})\boldsymbol{f} - \frac{1}{2}\boldsymbol{f}^\top K^{-1}\boldsymbol{f}\right] \\ \propto& \exp\left[\left(g(\boldsymbol{y}) + \beta(\boldsymbol{y}) \circ \boldsymbol{\omega}\right)^\top \boldsymbol{f} - \boldsymbol{f}^\top\left[\operatorname{diag}(\gamma(\boldsymbol{y}) \circ \boldsymbol{\omega}) + \frac{1}{2}K^{-1}\right]\boldsymbol{f}\right]. \end{aligned}$$

We get immediately a multivariate normal distribution with $-\frac{1}{2}\Sigma^{-1} = -\operatorname{diag}(\gamma(\boldsymbol{y}) \circ \boldsymbol{\omega}) + \frac{1}{2}K^{-1}$ and $\Sigma^{-1}\mu = g(\boldsymbol{y}) + (\beta(\boldsymbol{y}) \circ \boldsymbol{\omega})$. Which corresponds to the result shown in equation (11).

For the augmented variable ω_i:

$$\begin{aligned} p(\omega_i|y_i, f_i) \propto& p(y_i|f_i, \omega_i)p(\omega_i) \\ \propto& \exp\left(-\|h(y_i, f_i)\|_2^2\omega_i\right)\pi_\varphi(\omega_i|0) \\ =& \pi_\varphi(\omega_i|\|h(y_i, f_i\|_2). \end{aligned}$$

Note that the equation 9 gives the normalization constant directly $\varphi(\|h(y_i, f_i)\|_2^2)$ directly. QED.

A.2 Computation of the moments and cumulants for the augmentation variable

Given the general class of distribution $\pi_\varphi(\omega|c)$ described in Section 3.1, moments and cumulants can be easily computed: The k-th moment of a distribution can be computed by taking the k-th derivative of the moment generating function (equivalent to a negative Laplace transform) at $t = 0$. For example for the first moment:

$$\begin{aligned} \mathbb{E}_{\pi_\varphi(\omega|c)}\left[\omega\right] =& \left.\frac{d\mathcal{L}\left\{\pi_\varphi(\omega|c)\right\}(-t)}{dt}\right|_{t=0} \\ =& \left.\frac{d}{dt}\left[\mathcal{L}\left[\frac{e^{-c^2\omega}\pi_\varphi(\omega|0)}{\varphi(c^2)}\right](-t)\right]\right|_{t=0} \\ =& -\frac{1}{\varphi(c^2)}\left.\frac{d}{dt}\left[\mathcal{L}\left[\pi_\varphi(\omega|b, 0)\right](t + c^2)\right]\right|_{t=0} \\ =& -\frac{1}{\varphi(c^2)}\left.\frac{d\varphi\left(t + c^2\right)}{dt}\right|_{t=0} \\ =& -\left.\frac{d\log\varphi(t)}{dt}\right|_{t=c^2} \\ =& -\frac{\varphi'(c^2)}{\varphi(c^2)} = \overline{\omega} \end{aligned}$$

More generally the k-th moment m_k is defined as :

$$m_k = (-1)^k \frac{1}{\varphi(c^2)} \left. \frac{d^k \varphi(t)}{dt^k} \right|_{c^2}$$

And the cumulants κ_k are computed using the cumulant generating function (log of the moment generating function)

$$\kappa_k = (-1)^k \left. \frac{d^k \log \varphi(t)}{dt^k} \right|_{t=c^2}$$

A.3 Algorithm for the sparse case

Algorithm 3 Augmented Stochastic Variational Inference

Input: Data $(\mathbf{X}, \boldsymbol{y})$, GP model $p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{u})$, kernel k
Output: Approximate posterior $q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{m}, \boldsymbol{S})$
Find inducing points inputs Z via k-means
Compute kernel matrices : K_Z, $\kappa = K_{XZ} K_Z^{-1}$
for iteration $t = 1, 2, \ldots,$ **do**
  `# Local updates:`
  Sample minibatch $\mathcal{B} \subseteq \{1, \ldots, n\}$
  for $i \in \mathcal{B}$ **do**
    $c_i = \sqrt{\mathbb{E}_{q(f)}\left[h(f_i, y_i)^2\right]}$
    $\overline{\omega}_i = \mathbb{E}_{q(\omega_i)}\left[\omega_i\right] = -\varphi'(c_i^2)/\varphi(c_i^2)$
  end for
  `# Natural gradient updates (CAVI):`
  $\widetilde{\boldsymbol{S}} = \left(\kappa^\top \operatorname{diag}\left(2\overline{\boldsymbol{\omega}} \circ \gamma(\boldsymbol{y})\right) \kappa + K_Z^{-1}\right)^{-1}$
  $\widetilde{\boldsymbol{m}} = \widetilde{\boldsymbol{S}} \left(K_Z^{-1} \mu_0 + \kappa^\top \left(g(\boldsymbol{y}) + \overline{\boldsymbol{\omega}} \circ \beta(\boldsymbol{y})\right)\right)$
  $\{\boldsymbol{m}, \boldsymbol{S}\} \leftarrow (1 - \rho^{(t)})\{\boldsymbol{m}, \boldsymbol{S}\} + \rho^{(t)}\{\widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{S}}\}$
end for

$\rho^{(t)}$ is an arbitrary learning rate respecting the Robbins-Monroe condition.

A.4 ELBO Analysis

A.4.1 Full ELBO

$$\begin{aligned}\text{ELBO} = &\sum_{i=1}^{N} \mathbb{E}_{q(f_i, \omega_i)}\left[\log p(y_i|f_i, \omega_i)\right] \\ &- \mathrm{KL}[q(f)||p(f)] - \sum_{i=1}^{N} \mathrm{KL}[q(\omega_i)||p(\omega_i)]\end{aligned}$$

$$\begin{aligned}\mathbb{E}_q\left[\log p(y_i|f_i, \omega_i, \theta)\right] &= \log C(\theta) + g(y_i, \theta)\mathbb{E}_{q(f)}\left[f\right] - \mathbb{E}_{q(f)}\left[h(f_i, y_i)^2\right]\mathbb{E}_{q(\omega_i)}\left[\omega_i\right] \\ &= \log C(\theta) + g(y_i, \theta) m_i - \left(\alpha(y_i) - \beta(y_i) m_i + \gamma(y_i)\left(m_i^2 + S_{ii}\right)\right)\overline{\omega}_i \\ \mathrm{KL}[q(f)||p(f)] &= \frac{1}{2}\left[\log \frac{|K|}{|\boldsymbol{S}|} - N + \operatorname{tr}(K^{-1}\boldsymbol{S}) + (\boldsymbol{\mu}_0 - \boldsymbol{m})^\top K^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{m})\right] \\ \mathrm{KL}[q(\omega_i)||p(\omega_i)] &= -\mathbb{E}_{q(\omega_i)}\left[c_i^2 \omega_i\right] - \log \varphi(c_i^2) = -c_i^2 \overline{\omega}_i - \log \varphi(c_i^2)\end{aligned}$$

Note that we can take the derivatives of the ELBO and set them to 0 to recover exactly the updates in algorithm 1.

A.4.2 Analysis of the optima

By setting c_i^2 as a function of $\boldsymbol{m}$ and $\boldsymbol{S}$ (and setting $\boldsymbol{\mu}_0$ to 0 for simplicity) we can get an ELBO only depending of the variational parameters of $\boldsymbol{f}$.

$$\text{ELBO}(\boldsymbol{m}, \boldsymbol{S}) = C + g^\top \boldsymbol{m} + \frac{1}{2}\left(\underbrace{\log|\boldsymbol{S}| - \text{tr}(K^{-1}\boldsymbol{S}) - \boldsymbol{m}^\top K^{-1}\boldsymbol{m}}_{\text{ELBO}_1}\right) + \sum_i \underbrace{\log\varphi(m_i^2 + S_{ii})}_{\text{ELBO}_2}$$

It is easy to show that ELBO_1 is jointly concave in $\boldsymbol{m}$ and $\boldsymbol{S}$ with a short matrix analysis. However ELBO_2 is more complex : $m_i^2 + S_{ii}$ is jointly convex in $\boldsymbol{m}$ and $\boldsymbol{S}$, $\phi(r)$ is by definition convex as well, however $\phi(m_i^2 + S_{ii})$ is neither jointly convex or concave in $\boldsymbol{m}$ and $\boldsymbol{S}$. It is therefore impossible to guarantee that there is a global optima, however the CAVI updates guarantee us a local optima.

A.4.3 ELBO Gap

For a fixed $q(f)$ we can compare the ELBO of the original model $\mathcal{L}_{std}(q(f))$ and the augmented model $\mathcal{L}_{aug}(q(f)q(\omega))$. It is then straightforward to compute the difference between the two :

$$\begin{aligned}
\Delta\mathcal{L} =& \mathcal{L}_{std}(q(f)) - \mathcal{L}_{aug}(q(f)q(\omega)) \\
=& \mathbb{E}_{q(f)}\left[\log p(y,f) - \log q(f) - \mathbb{E}_{q(\omega)}\left[p(y,f,\omega) - \log q(f)q(\omega)\right]\right] \\
=& \mathbb{E}_{q(f)q(\omega)}\left[-\log\frac{p(y,f,\omega)}{p(y,f)} + \log q(\omega)\right] \\
=& \mathbb{E}_{q(f)q(\omega)}\left[-\log p(\omega|y,f) + \log q(\omega)\right] \\
=& \mathbb{E}_{q(\omega)}\left[\log q(\omega) - \mathbb{E}_{q(f)}\left[\log p(\omega|y,f)\right]\right] \\
=& -c^2\mathbb{E}_{q(\omega)}\left[\omega\right] + \mathbb{E}_{q(\omega)}\left[\log \text{PG}(\omega|1,0)\right] - \log\varphi(c^2) \\
&+ \mathbb{E}_{q(f)}\left[f^2\right]\mathbb{E}_{q(\omega)}\left[\omega\right] - \mathbb{E}_{q(\omega)}\left[\log \text{PG}(\omega|1,0)\right] + \mathbb{E}_{q(f)}\left[\log\varphi(f^2)\right] \\
=& -c^2 m - \log\varphi(c^2) + \mathbb{E}_{q(f)}\left[f^2\right] m + \mathbb{E}_{q(f)}\left[\log\varphi(f^2)\right]
\end{aligned}$$

Replacing with the optimal $q^*(\omega) = \frac{e^{-c^2\omega}p(\omega)}{\varphi(c^2)}$ with $c^2 = \mathbb{E}_{q(f)}\left[f^2\right]$

$$\Delta\mathcal{L}^* = -\log\varphi(c^2) + \mathbb{E}_{q(f)}\left[\log\varphi(f^2)\right]$$

A.4.4 Sparse ELBO

When using the inducing points approach the ELBO becomes:

$$\begin{aligned}
\text{ELBO} =& \sum_{i=1}^N \mathbb{E}_{q(f_i,u_i,\omega_i)}\left[\log p(y_i|f_i,u_i,\omega_i)\right] \\
&- \text{KL}[q(u)||p(u)] - \sum_{i=1}^N \text{KL}[q(\omega_i)||p(\omega_i)]
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}_q\left[\log p(y_i|f_i,\omega_i,\theta)\right] &= \log C(\theta) + g(y_i,\theta)\mathbb{E}_{q(f,u)}\left[f\right] - \mathbb{E}_{q(f,u)}\left[h(f_i,y_i)^2\right]\mathbb{E}_{q(\omega_i)}\left[\omega_i\right] \\
&= \log C(\theta) + g(y_i,\theta)(\kappa^\top \boldsymbol{m})_i - \left(\alpha(y_i) - \beta(y_i)(\kappa^\top \boldsymbol{m})_i + \gamma(y_i)\left((\kappa^\top \boldsymbol{m})_i^2 + (\kappa^\top \boldsymbol{S}\kappa)_{ii}\right)\right)\overline{\omega}_i \\
\mathrm{KL}[q(f)||p(f)] &= \frac{1}{2}\left[\log\frac{|K|}{|\boldsymbol{S}|} - N + \mathrm{tr}(K^{-1}\boldsymbol{S}) + (\boldsymbol{\mu}_0 - \boldsymbol{m})^\top K^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{m})\right] \\
\mathrm{KL}[q(\omega_i)||p(\omega_i)] &= -\mathbb{E}_{q(\omega_i)}\left[c_i^2\omega_i\right] - \log\varphi(c_i^2) = -c_i^2\overline{\omega}_i - \log\varphi(c_i^2)
\end{aligned}
$$

A.5 Proof of equivalence between Jaakkola bound and data augmentation

Jaakkola and Jordan (2000) proposed an approach purely based on optimization. They are assuming $\log p(y|f)$ contains a part convex in f^2: $\log p(y|f) = \log p_{convex}(f) + \log p_{non-convex}(f,y)$. Using convexity properties they are creating a bound with a Taylor expansion to the first order around an additional variable c^2:

$$
\log p_c(f) \geq \log p_c(c) + \frac{d\log p_c(c)}{dc^2}(f^2 - c^2)
$$

Putting it back in the full ELBO, they are now getting a quadratic part in f, analytically differentiable, and they just need to optimize the additional variables $\{c_i\}$. Merkle (2014) shows that any completely monotone function is log-convex, i.e. $\log\varphi(r)$ is convex. Therefore we can replace $\log p_c(c)$ by $\log\varphi(r)$ to recover our model in the context of variational inference. Note that the converse does not hold, therefore the complete monotonicity is a stronger assumption.

A.6 Likelihoods used for the experiments

We detail all likelihoods used for the experiments and their formulation as in equation (4).

Laplace Likelihood : $\mathrm{Laplace}(y|f,\beta) = \frac{1}{2\beta}\exp\left(-\frac{|f-y|}{\beta}\right)$

Logistic Likelihood : $p(y|f) = \sigma(yf) = \frac{e^{yf/2}}{2\cosh(|f|/2)}$

Student-T Likelihood : $p(y|f) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu}}\left(1 + \frac{(y-f)^2}{\nu}\right)^{-(\nu+1)/2}$

Matern 3/2 Likelihood : $p(y|f) = \frac{4\rho}{\sqrt{3}}\left(1 + \frac{\sqrt{3(y-f)^2}}{\rho}\right)\exp\left(-\frac{\sqrt{3(y-f)^2}}{\rho}\right)$

Likelihood	$C(\theta)$	$g(y,\theta)$	$\|\|h(y,f,\theta)^2\|\|_2^2$	$\alpha(y)$	$\beta(y)$	$\gamma(y)$	$\varphi(r)$
Laplace	$(2\beta)^{-1}$	0	$(y-f)^2$	y^2	$2y$	1	$e^{-\sqrt{r}/\beta}$
Logistic	2^{-1}	$y/2$	f^2	0	0	1	$\cosh^{-1}(\sqrt{r}/2)$
Student-T	$\Gamma((\nu+1)/2)/(\Gamma(\nu)\sqrt{\pi\nu})$	0	$(y-f)^2$	y^2	$2y$	1	$(1+\frac{r}{\nu})^{-(\nu+1)/2}$
Matern 3/2	$4\rho/\sqrt{3}$	0	$(y-f)^2$	y^2	$2y$	1	$(1+\frac{\sqrt{3r}}{\rho})e^{-\sqrt{3r}/\rho}$

A.7 Extra figures

A.7.1 Autocorrelation plots

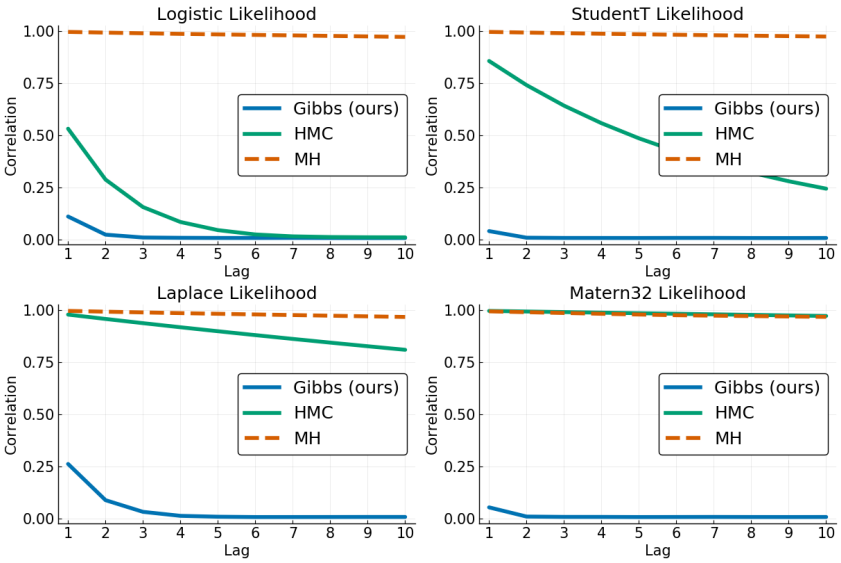


Figure 4. Auto-correlation plots for differents with lags from 1 to 10

A.7.2 HMC Results

ϵ/n_{step}		1	2	5	10
0.01	Time/Sample (s)	0.037	0.045	0.077	0.133
	Lag 1	0.999	0.993	0.978	0.963
	Gelman	3.14	1.02	1.00	2.05
0.05	Time/Sample (s)	0.036	0.046	0.080	0.12
	Lag 1	0.999	0.998	**0.931**	0.948
	Gelman	1.72	1.18	1.01	3.25
0.1	Time/Sample (s)	0.033	0.042	0.073	0.13
	Lag 1	0.997	0.996	0.998	0.994
	Gelman	1.11	1.04	1.27	2.71

Table 3. HMC results for the Laplace likelihood

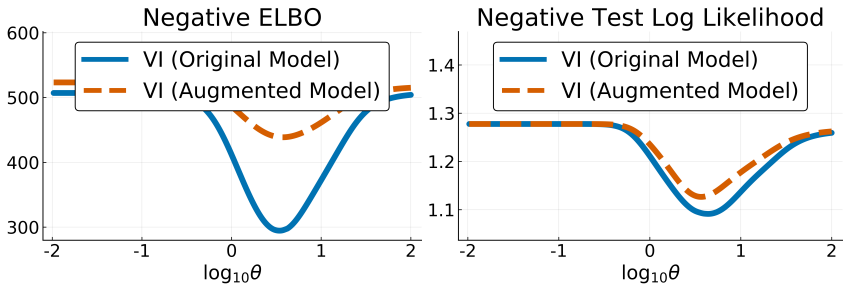
ϵ/n_{step}		1	2	5	10
0.01	Time/Sample (s)	0.675	0.110	0.177	0.251
	Lag 1	0.999	0.999	0.997	0.993
	Gelman	3.14	1.74	1.11	1.02
0.05	Time/Sample (s)	0.148	0.192	0.336	0.573
	Lag 1	0.997	0.993	0.962	**0.857**
	Gelman	1.10	1.02	1.00	1.00
0.1	Time/Sample (s)	0.142	0.193	0.337	NA
	Lag 1	0.993	0.976	0.864	NA
	Gelman	1.03	1.01	1.00	NA

Table 4. HMC results for the Student-T likelihood

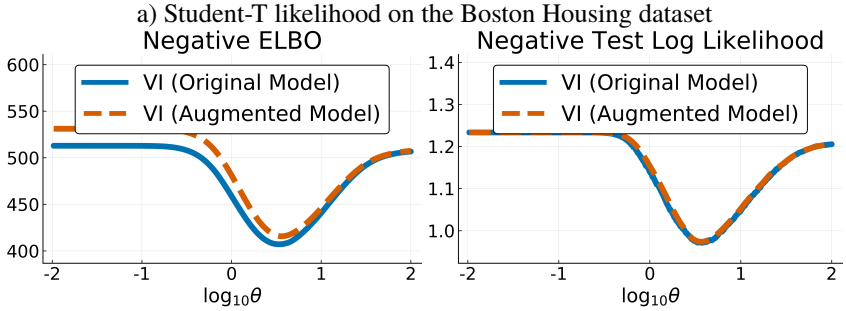
ϵ/n_{step}		1	2	5	10
0.01	Time/Sample (s)	0.009	0.013	0.021	0.041
	Lag 1	0.999	0.999	0.998	0.994
	Gelman	3.19	1.68	1.12	1.02
0.05	Time/Sample (s)	0.011	0.014	0.025	0.41
	Lag 1	0.998	0.994	0.968	0.871
	Gelman	1.11	1.03	1.00	1.00
0.1	Time/Sample (s)	0.011	0.014	0.024	0.048
	Lag 1	0.994	0.979	0.875	**0.532**
	Gelman	1.02	1.01	1.00	1.00

Table 5. HMC Results for the Logistic likelihood

A.7.3 ELBO difference



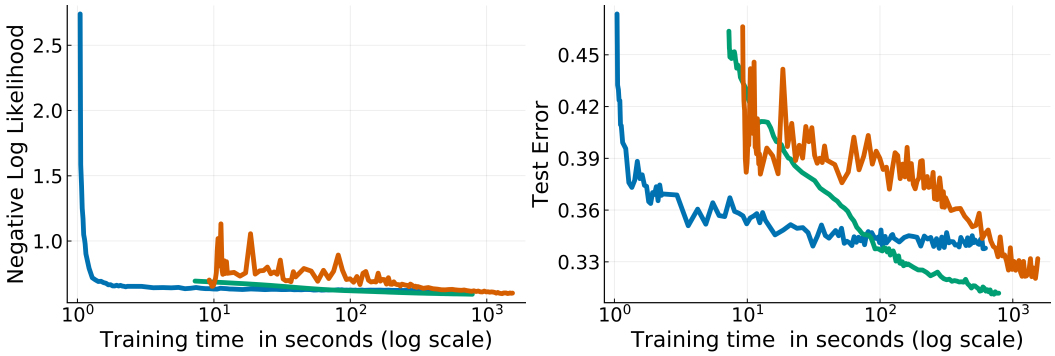
a) Student-T likelihood on the Boston Housing dataset



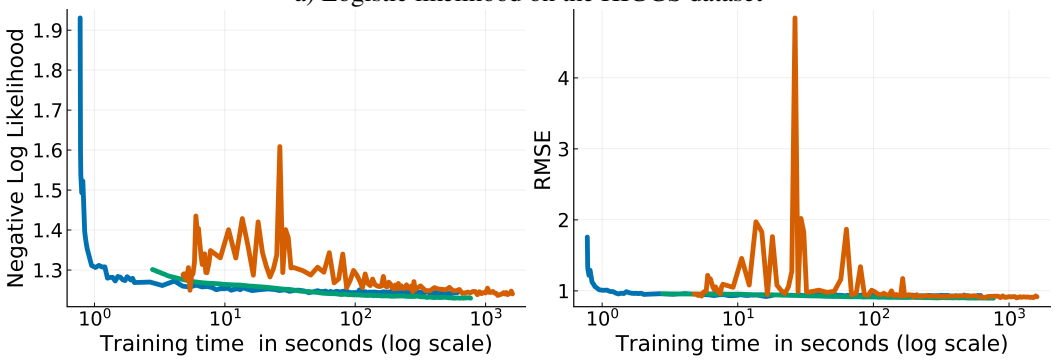
b) Laplace likelihood on the Boston Housing dataset

Figure 5. Converged negative ELBO and averaged negative log-likelihood on a held-out dataset in function of the RBF kernel lengthscale, training VI with and without augmentation.

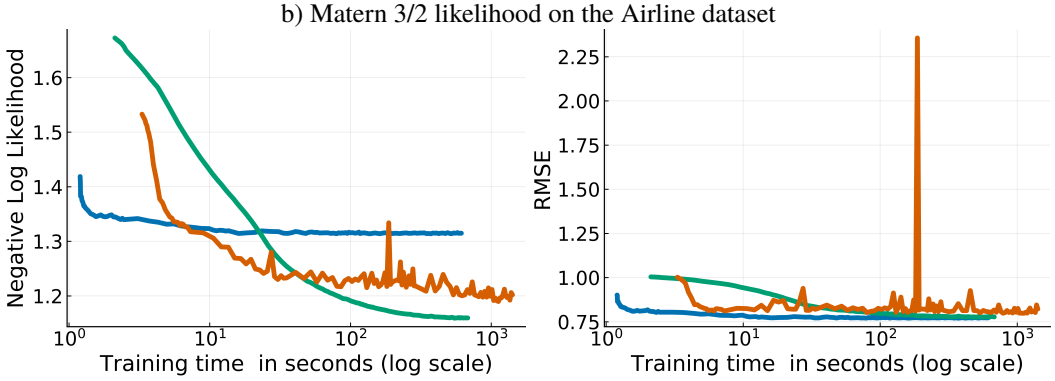
A.7.4 Convergence speed



a) Logistic likelihood on the HIGGS dataset



b) Matern 3/2 likelihood on the Airline dataset



c) Student-T likelihood on the Protein dataset

Figure 6. Supplementary convergence plots