Supplementary material for: Doubly Stochastic Variational Bayes for non-Conjugate Inference

A Proof for the concavity of the bound when $\log g(\boldsymbol{\theta})$ is concave

Proposition 1. Assume a continuous probability density function $\phi(\mathbf{z})$ in $\mathbb{R}^D$ and a joint density model function $g(\boldsymbol{\theta})$ which is log concave with respect to $\boldsymbol{\theta} \in \mathbb{R}^D$. The variational lower bound

$$\mathcal{F}(\boldsymbol{\mu}, C) = \int_{\mathbb{R}^D} q(\boldsymbol{\theta}|\boldsymbol{\mu}, C) \log \frac{g(\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{\mu}, C)} d\boldsymbol{\theta},$$

where $q(\boldsymbol{\theta}|\boldsymbol{\mu}, C) = \frac{1}{|C|}\phi\left(C^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)$, is concave with respect $(\boldsymbol{\mu}, C)$.

Proof: By applying the transformation $\mathbf{z} = C^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})$ where C is taken to be a positive definite matrix, $\mathcal{F}(\boldsymbol{\mu}, C)$ can be written as

$$\mathcal{F}(\boldsymbol{\mu}, C) = \int_{\mathbb{R}^D} \phi(\mathbf{z}) \left[\log g(C\mathbf{z} + \boldsymbol{\mu})\right] d\mathbf{z} + \log|C| + \mathcal{H}(\phi(\mathbf{z})),$$

where $\mathcal{H}(\phi(\mathbf{z}))$ is constant with respect to $(\boldsymbol{\mu}, C)$. Since C is positive definite, $\log|C|$ is concave. Further, if $\log g(C\mathbf{z}+\boldsymbol{\mu})$ is concave with respect to $(\boldsymbol{\mu}, C)$, then $\int_{\mathbb{R}^D} \phi(\mathbf{z}) \left[\log g(C\mathbf{z} + \boldsymbol{\mu})\right] d\mathbf{z}$ would also be concave since it would be a non-negative linear combination of concave functions and for the same reason $\mathcal{F}(\boldsymbol{\mu}, C)$ would overall be concave. Therefore, what remains to show is that $\log g(C\mathbf{z} + \boldsymbol{\mu})$ is concave with respect to $(\boldsymbol{\mu}, C)$. It holds that for $\alpha \in [0, 1]$, $\bar{\alpha} = 1 - \alpha$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^D$

$$\log g(\alpha\boldsymbol{\theta}_1 + \bar{\alpha}\boldsymbol{\theta}_2) \geq \alpha \log g(\boldsymbol{\theta}_1) + \bar{\alpha} \log g(\boldsymbol{\theta}_2),$$

because $\log g(\boldsymbol{\theta})$ is concave with respect to $\boldsymbol{\theta}$. We want to show that

$$E(C, \boldsymbol{\mu}) = \log g(C\mathbf{z} + \boldsymbol{\mu})$$

is concave w.r.t. $(C, \boldsymbol{\mu})$. We have

$$\begin{aligned} E(\alpha C_1 + \bar{\alpha} C_2, \alpha\boldsymbol{\mu} + \bar{\alpha}\boldsymbol{\mu}) &= \log g\left(\alpha(C_1\mathbf{z} + \boldsymbol{\mu}_1) + \bar{\alpha}(C_2\mathbf{z} + \boldsymbol{\mu}_2)\right) \\ &\geq \alpha \log g(C_1\mathbf{z} + \boldsymbol{\mu}_1) + \bar{\alpha} \log g(C_2\mathbf{z} + \boldsymbol{\mu}_2) \\ &= \alpha E(C_1, \boldsymbol{\mu}_1) + \bar{\alpha} E(C_2, \boldsymbol{\mu}_2) \end{aligned} \tag{1}$$

which means that $\log g(C\mathbf{z} + \boldsymbol{\mu})$ is jointly concave with respect to the variational parameters $(\boldsymbol{\mu}, C)$.

B DSVI for variable selection

Pseudo-code for the application of DSVI for variable selection is given in Algorithm 2 below. The instantaneous value for the lower bound (a rolling-window average of such values, of size 200, is displayed in Figure 3 in the main paper and Figure 1 below) is defined as a single-sample Monte Carlo estimate of the exact bound. I.e. at the tth iteration of the algorithm the instantaneous value of the lower bound is

$$\mathcal{F}^{(t)} = \log g(\boldsymbol{\theta}^{(t)}) + \log|C^{(t)}| + \mathcal{H}_\phi, \quad \boldsymbol{\theta}^{(t)} \sim q(\boldsymbol{\theta}|\boldsymbol{\mu}, C).$$

This corresponds to the general DSVI algorithm while for the DSVI-ARD case the instantaneous bound is defined analogously.

C Further information for the Gaussian process hyperparameter inference experiments

For the GP regression experiments the joint probability density $g(\boldsymbol{\theta})$ is written in the form

$$g(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma^2 I) p(\boldsymbol{\theta}),$$

Algorithm 2 Doubly stochastic variational inference for variable selection (DSVI-ARD)

Input: ϕ, $\mathbf{y}$, $\boldsymbol{\theta}$, $\nabla \log \widetilde{g}$.
Initialise $\boldsymbol{\mu}^{(0)}, \mathbf{c}^{(0)}$, $t = 0$.
repeat
$t = t + 1$;
$\mathbf{z} \sim \phi(\mathbf{z})$;
$\boldsymbol{\theta}^{(t-1)} = \mathbf{c}^{(t-1)} \circ \mathbf{z} + \boldsymbol{\mu}^{(t-1)}$;
$\mu_d^{(t)} = \mu_d^{(t-1)} + \rho_t \left(\frac{\partial \log \widetilde{g}(\boldsymbol{\theta}^{(t-1)})}{\partial \theta_d} - \frac{\mu_d^{(t-1)}}{(c_d^{(t-1)})^2 + (\mu_d^{(t-1)})^2} \right), \quad d = 1, \ldots, D;$
$c_d^{(t)} = c_d^{(t-1)} + \rho_t \left(\frac{\partial \log \widetilde{g}(\boldsymbol{\theta}^{(t-1)})}{\partial \theta_d} z_d + \frac{1}{c_d^{(t-1)}} - \frac{c_d^{(t-1)}}{(c_d^{(t-1)})^2 + (\mu_d^{(t-1)})^2} \right), \quad d = 1, \ldots, D;$
until convergence criterion is met.

where $\mathbf{K}$ is the covariance matrix defined from the GP prior, $\boldsymbol{\theta} = (\log(\ell_1^2), \ldots . \log(\ell_D^2), \log(\sigma_f^2), \log(\sigma^2))$ and $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, 10I)$ is the Gaussian prior. Notice that $\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma^2 I)$ is simply the marginal likelihood of the GP regression model.

Furthermore, DSVI is applied by using a full scale matrix C so that a fully dependent Gaussian approximation is fitted to the exact posterior over the hyperparameters. The learning rate sequences and annealing schedule when applying DSVI to all GP hyperparameter inference problems were chosen as follows. The learning rate ρ_t is initialised to $\rho_0 = 0.5/\#\text{training examples}$ and scaled every 1000 iterations by a factor of 0.95. This learning rate is used to update $\boldsymbol{\mu}$, whereas $0.1\rho_t$ was used to update C. A total of 20000 iterations was considered for all problems (i.e. 20 stages in the annealing schedule where in each stage the learning rate remains constant).

Next, we provide all plots for the three GP regression experiments. Firstly, Figure 1 shows the evolution of rolling-averages of instantaneous lower bounds computed as described previously. Then, Figures 2, 3 and 4 display the complete set of the marginal posterior distributions for all hyperparameters in the three datasets.

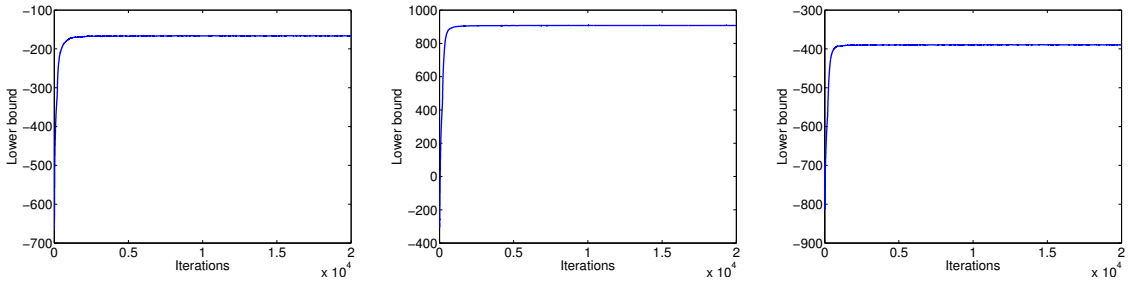

Figure 1: Rolling-averages (over a window of previous 200 iterations) of the instantaneous lower bound values for `Boston` (left), `Bodyfat` (middle) and `Pendulum` (right) datasets.

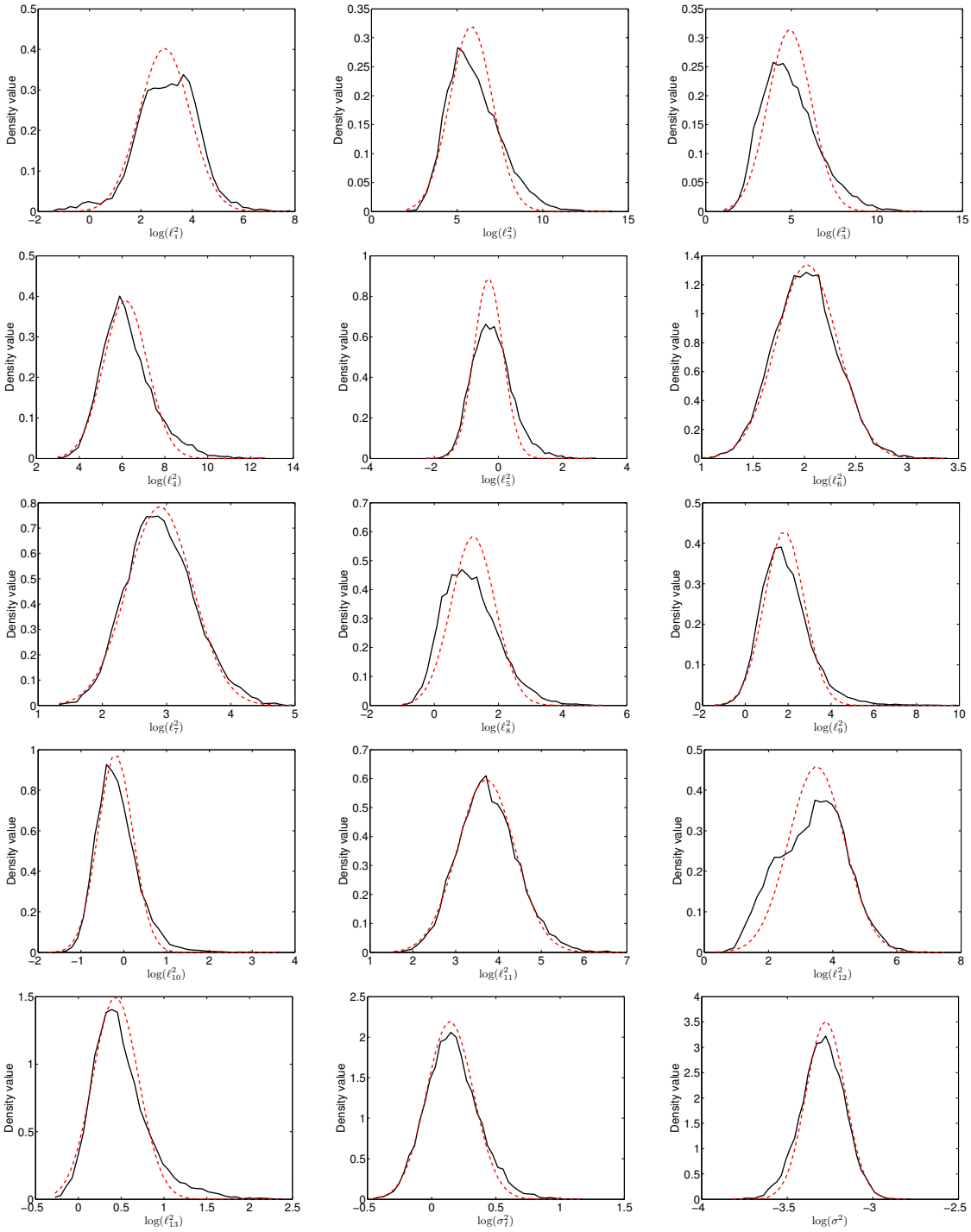

Figure 2: All marginal posterior distributions for all hyperparameters in `Boston` dataset. The black solid lines show the ground-truth empirical estimates obtained by a very long run of MCMC. The red dashed lines show the Gaussian marginals found by stochastic variational inference.

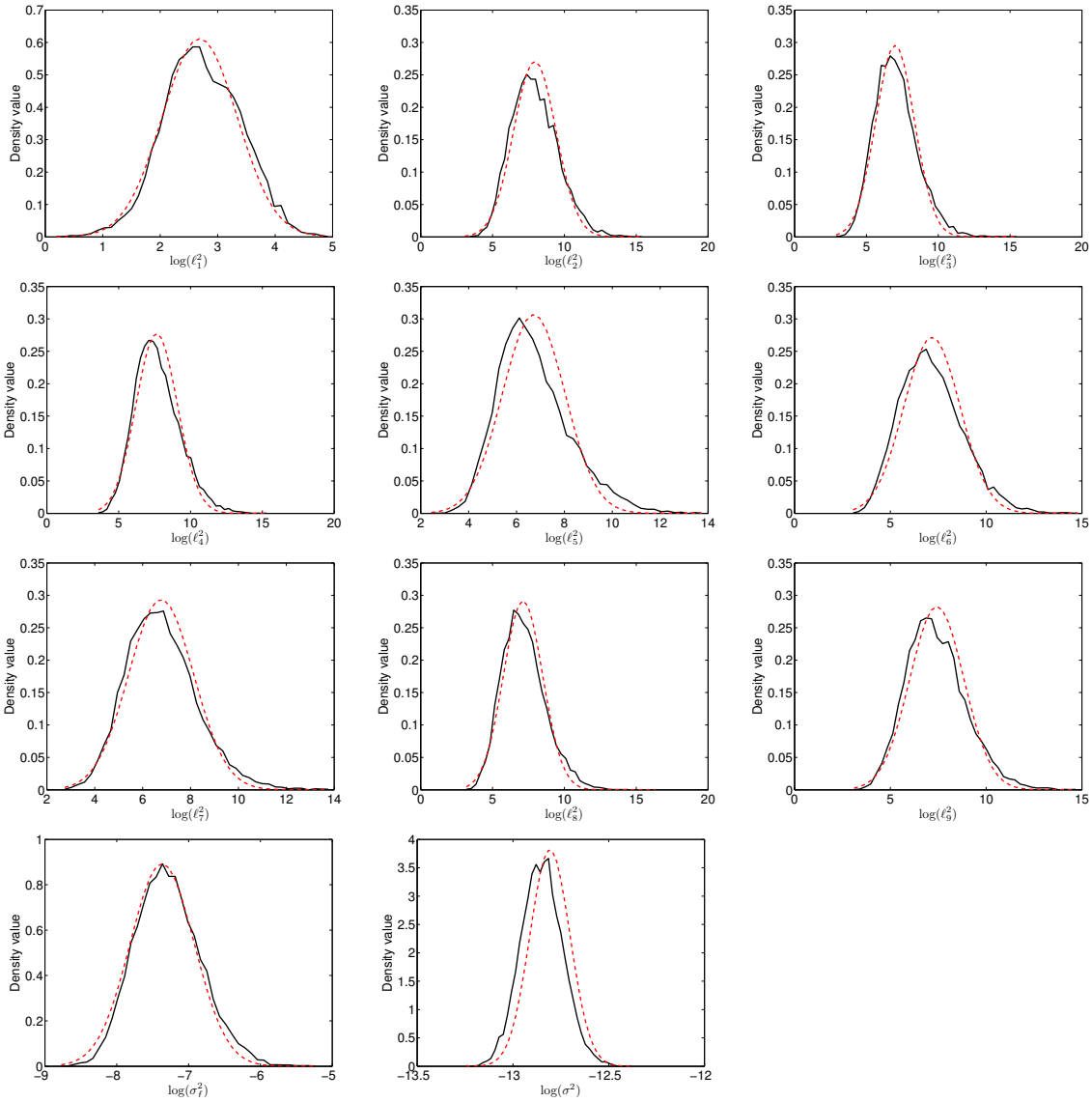

Figure 3: All marginal posterior distributions for all hyperparameters in `Bodyfat` dataset. The black solid lines show the ground-truth empirical estimates obtained by a very long run of MCMC. The red dashed lines show the Gaussian marginals found by stochastic variational inference.

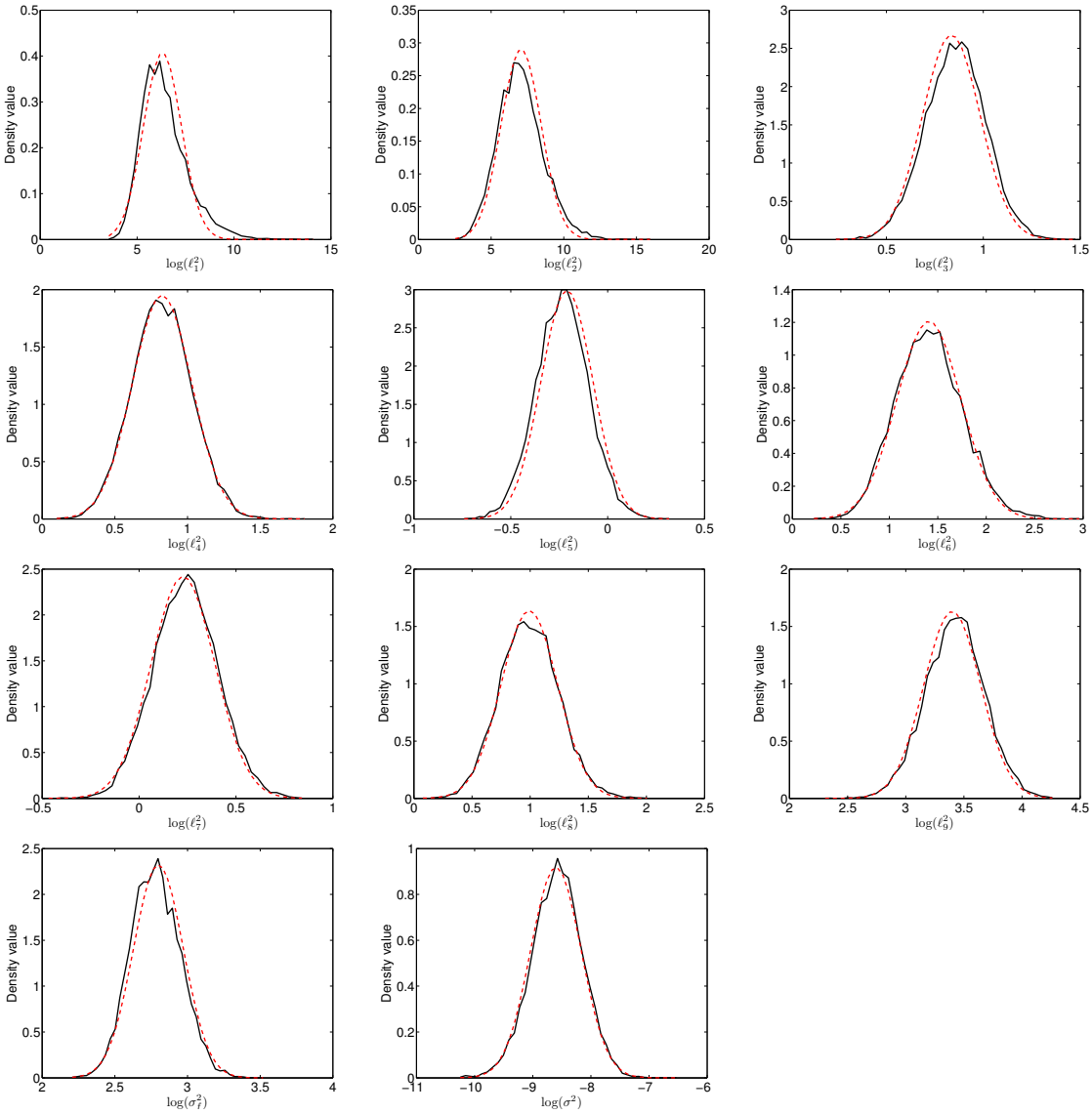

Figure 4: All marginal posterior distributions for all hyperparameters in `Pendulum` dataset. The black solid lines show the ground-truth empirical estimates obtained by a very long run of MCMC. The red dashed lines show the Gaussian marginals found by stochastic variational inference.