Appendix

A Tightness of Variational Bound in the Limit $M \to \infty$,

In the following, we show for the one-dimensional case that for a transformation function built from Bernstein polynomials of order M there exists a solution so that $\mathrm{KL}(q_{\lambda,M}(\theta) \parallel p(\theta \mid D))$ in (3) asymptotically vanishes with $1/M$.

We assume that the posterior $p(\theta \mid D)$ is continuous in the compact interval $[\theta_a, \theta_b]$ with CDF $F_\theta(\theta)$ and that the first three derivatives of $F_\theta(\theta)$ exists. Without loss of generality, we set $\theta_a = 0$ and $\theta_b = 1$. This allows us to construct a transformation $h : \theta \to Z$ transforming between θ with CDF F_θ and a random variable Z with CDF F_Z by requiring $F_Z(h(\theta)) = F_\theta(\theta)$ via

$$h(\theta) = F_Z^{-1}(F_\theta(\theta)). \tag{S1}$$

We assume that F_Z has continuous derivatives up through order 3, then from (S1), it follows that $h(\theta)$ has continuous derivatives up to order 3. Further, we require that the number of roots of $h''(\theta)$ and $h'''(\theta)$ are countable. In addition, we require that the density $p_Z(z) > 0$. We approximate the posterior density $p(\theta \mid D)$ by $q_\lambda(\theta)$ by approximating $h(\theta)$ with a Bernstein polynomial of order M given by:

$$B_M(\theta) = \sum_{m=0}^{M} h\left(\frac{m}{M}\right) \binom{M}{m} \theta^m (1-\theta)^{M-m} \tag{S2}$$

Note that we fixed the coefficients of the transformation here. This is valid since we just want to prove that a particular solution exists for that (3) asymptotically vanishes with $1/M$. As shown in [32], the Bernstein approximation "is at least as smooth" as $h(\theta)$, in our case that at least the first 3 derivatives of $B_M(\theta)$ exists and converge to the corresponding derivatives of $h(\theta)$. In a first step, we upper bound the KL-divergence to that part of $\theta' \subset \theta$ where $\log(\frac{q_\lambda(\theta)}{p(\theta|D)}) > 0$ since $q_\lambda(\theta) \geq 0$, the following approximation hold

$$\mathrm{KL}(q_\lambda(\theta) \parallel p(\theta \mid D)) = \int_0^1 q_\lambda(\theta) \log\left(\frac{q_\lambda(\theta)}{p(\theta \mid D)}\right) d\theta \leq \int_{\theta'} q_\lambda(\theta) \log\left(\frac{q_\lambda(\theta)}{p(\theta \mid D)}\right) d\theta$$

Using the change of variable formula, we can express the densities of the posterior as

$p(\theta \mid D) = p_Z(h(\theta)) \cdot |h'(\theta)|$ and its approximation as $q_\lambda(\theta) = p_Z(B_M(\theta)) \cdot |B'_M(\theta)|$, the dash indicates a derivative w.r.t. θ. Hence,

the KL-Divergence is bounded by:

$$\mathrm{KL}(q_\lambda(\theta) \parallel p(\theta \mid D)) \leq \int_{\theta'} q_\lambda(\theta) \log\left(\frac{p_Z\left(B_M(\theta)\right) \cdot |B'_M(\theta)|}{p_Z\left(h(\theta)\right) \cdot \mid h'(\theta) \mid}\right) d\theta \tag{S3}$$

The transformation function $h(\theta)$ is either strictly monotonic increasing or decreasing, w.l.o.g. we assume that $h'(\theta) > 0$. This leads to ordered coefficients $h(m/M)$ and so $B'_M(\theta) > 0$.

Since the density $p_Z(z)$ and the derivative $B'_M(\theta)$ of Bernstein approximation are finite, so is $q_\lambda(\theta) = p_Z\left(B_M(\theta)\right) \cdot \mid B'_M(\theta) \mid$. This allows to upper bound the integral in (S3) with $C = \sup(q_\lambda(\theta))$ in the range $\theta \in [0, 1]$ by:

$$\mathrm{KL}(q_\lambda(\theta) \parallel p(\theta \mid D)) \leq C \cdot \underbrace{\int_{\theta'} \log\left(\frac{p_Z\left(B_M(\theta)\right)}{p_Z\left(h(\theta)\right)}\right) d\theta}_{=:I_1} + C \cdot \underbrace{\int_{\theta'} \log\left(\frac{B'_M(\theta)}{h'(\theta)}\right) d\theta}_{=:I_2}$$

With increasing M the term $\Delta = B_M(\theta) - h(\theta)$ gets arbitrarily small. We do a Taylor expansion $p_Z(B_M(\theta)) = p_Z(h(\theta) + \Delta) = p_Z(h(\theta)) + p'_Z(h(\theta))\Delta + \mathcal{O}(\Delta^2)$. With $\log(1 + x) \leq x$ the first integral I_1 can be approximated as

$$I_1 = \int_{\theta'} \log\left(1 + \frac{p'_Z\left(h(\theta)\right)}{p_Z\left(h(\theta)\right)}\Delta + \mathcal{O}(\Delta^2)\right) d\theta \leq \int_{\theta'} \left(\frac{p'_Z\left(h(\theta)\right)}{p_Z\left(h(\theta)\right)}\Delta + \mathcal{O}(\Delta^2)\right) d\theta$$

Since $p_Z(z) > 0$ we can set $C_2 = \sup(\frac{p'_Z(h(\theta))}{p_Z(h(\theta))})$ yielding:

$$I_1 \leq C_2 \cdot \int_{\theta'} \left(B_M(\theta) - h(\theta)\right) d\theta + \mathcal{O}(\Delta^2)$$

The Voronovskaya theorem (see [32]) states that $B_M(\theta) - h = \theta/(2M)h''(\theta) + o(M^{-1})$ for $h''(\theta) \neq 0$. Assuming that $h''(\theta) = 0$ only on a countable set of points, the asymptotic of I_1 is given by $1/M$.

Using $\log(x) \leq x - 1$, the second integral can be bounded as follows.

$$I_2 = \int_{\theta'} \log\left(\frac{B'_M(\theta)}{h'(\theta)}\right) d\theta \leq \int_{\theta'} \left(\frac{B'_M(\theta)}{h'(\theta)} - 1\right) d\theta = \int_{\theta'} \frac{1}{h'(\theta)} \left(B'_M(\theta) - h'(\theta)\right) d\theta$$

According to our assumptions $h'(\theta) > 0$ and so the supremum $C_3 = \sup(1/h'(\theta))$ exists.

$$I_2 \leq C_3 \int_{\theta'} \left(B'_M(\theta) - h'(\theta)\right) d\theta \leq \frac{1}{M} \int_{\theta'} \left(\theta/2h'''(\theta) + 1/2h''(\theta) + o(M^{-1})\right) d\theta$$

This shows that the KL-Divergence in (3) asymptotically converges to zero as $1/M$.

B Experimental Details of and Additional Results

In the following, we give additional details for the experiments. For the examples using MCMC simulations as a benchmark, the Stan-code can be found at https://github.com/tensorchiefs/bfvi_paper/tree/main/mcmc. If not otherwise stated, the training has been as described in the main text. If not given below, the used data can be found at: https://github.com/tensorchiefs/bfvi_paper/tree/main/data.

B.1 Bernoulli Example

The Data Generation Process

Two samples are drawn from a Bernoulli distribution.

The Actual Data

The realizations are $y_1 = 1$ and $y_2 = 2$.

Details of the Training Procedure

To demonstrate the possibility that the described solution converges in principle to the exact solution, we used a large number of MC samples $S = 2500$ and epochs=2500 for the training. Figure S1 shows the loss curve for different values of M ($M = 10$, 30, 50, and 100), along with Gauss-VI for comparison.

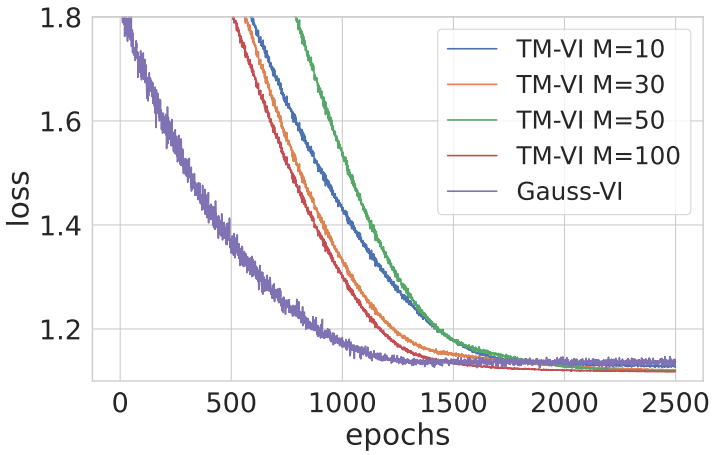

Fig. S1 The loss ELBO in the Bernoulli Example for $M = 10$, 30, 50, 100, and Gauss-VI.

B.1.1 Additional Results (Tightness of the ELBO)

For this example, we can estimate the KL divergence via

$$\mathrm{KL}(q_\lambda(\theta) \parallel p(\theta \mid D)) = \int q_\lambda(\theta) \log\left(\frac{q_\lambda(\theta)}{p(\theta \mid D)}\right) d\theta \approx \frac{1}{S} \sum_{\theta_s \sim q_\lambda} \log\left(\frac{q_\lambda(\theta_s)}{p(\theta_s \mid D)}\right)$$

The sum is calculated from the samples θ_s of the trained approximative distribution, which can be obtained via the trained transformation function. The likelihood $p(D \mid \theta_s)$ and prior $p(\theta_s)$ for those samples can be easily calculated and probabilities $q_\lambda(\theta_s)$ can be calculated via (6). In this simple case, the posterior can be calculated analytically to $p(\theta_s \mid D) = \mathrm{Be}_{(\alpha+\sum y_i, \beta+n-\sum y_i)}(\theta_s) = \mathrm{Be}_{(3.1,1.1)}(\theta_s)$.

In Fig. 2, we show a direct comparison of the posterior and its approximation of the dependence of the tightness of the ELBO on M, importantly increasing M and thus the flexibility does not deteriorate the approximation.

B.2 Cauchy Example

The Data Generation Process

Six samples are drawn from a mixture of two Cauchy distributions $y \sim \mathbf{Cauchy}(\xi_1 = -2.5, \gamma = 0.5) + \mathbf{Cauchy}(\xi_2 = 2.5, \gamma = 0.5)$.

The Actual Data

From the data generating process, we draw the following 6 examples $y = 1.2083935, -2.7329216, 4.1769943, 1.9710574, -4.2004027, -2.384988$.

Details of the Model

We use an unconditional Cauchy model $y \sim \mathbf{Cauchy}(\xi, \gamma = 0.5)$ were $\gamma = 0.5$ is given. Note that this model is misspecified and gives rise to a multimodal posterior. The location parameter ξ is assigned a standard normal prior $\xi \sim N(0, 1)$.

Details of the Training Procedure

To demonstrate the possibility that the described solution converges in principle to the exact solution, we used a rather large number of MC samples $S = 1000$ and epochs=1000 for the training. Figure S2 shows the loss curve (-ELBO) for different values of M ($M = 1$, $M = 10$, 30, and 50), along with Gauss-VI for comparison.

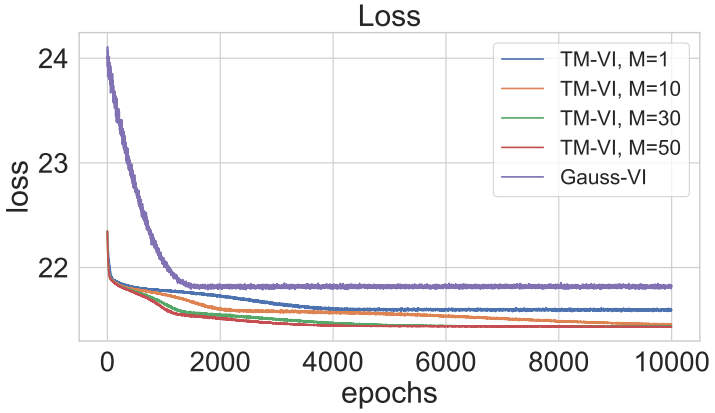

Fig. S2 The loss curve for $M = 1$, $M = 10$, 30, 50 and Gauss-VI for the Cauchy example.

B.2.1 Additional Results: Tightness of the ELBO

For this example, we cannot determine the posterior $p(\theta_s \mid D) = \frac{p(D|\theta_s)p(\theta_s)}{\int p(D|\theta_s)p(\theta_s)\, d\theta}$ **analytically anymore. However, we can rewrite (B.1.1) to:**

$$\begin{aligned} \mathrm{KL}(q_\lambda(\theta) \parallel p(\theta \mid D)) &\approx \frac{1}{S} \sum_{\theta_s \sim q_\lambda} \log\left(\frac{q_\lambda(\theta_s)}{p(\theta_s \mid D)}\right) \\ &= \frac{1}{S} \sum_{\theta_s \sim q_\lambda} \log\left(\frac{q_\lambda(\theta_s)}{p(D \mid \theta_s)p(\theta_s)}\right) + \underbrace{\log\left(\int p(D \mid \theta)p(\theta)d\theta\right)}_{=\log(P(D))} \end{aligned} \tag{S4}$$

Numerical integration obtains the log-evidence $\log(P(D)) = -21.43069$ **for this example.**

In Fig. S3, we show a direct comparison of posteriors. The dependence of the tightness of the ELBO on M **calculated in the lower part is calculated via (S4).**

B.2.2 Additional Results: Effect of Pre-Processing

In Fig. S3, the effect of different transformations before the Bernstein polynomials is studied.

There are different options to ensure that the input to BP fulfills the restriction $z \in [0, 1]$**. One option is to use a latent distribution with unrestricted. A random variable** z' **from that distribution is then "piped" through a sigmoid function** $\sigma : [-\infty, \infty] \to [0, 1]$**. Note that the total transformation** $f_{\mathrm{BP}} \circ \sigma$ **stays monotonically increasing. This construction allows us to choose the Standard-Normal for** $F_{Z'}$**. We have observed in low-dimensional examples a more efficient**

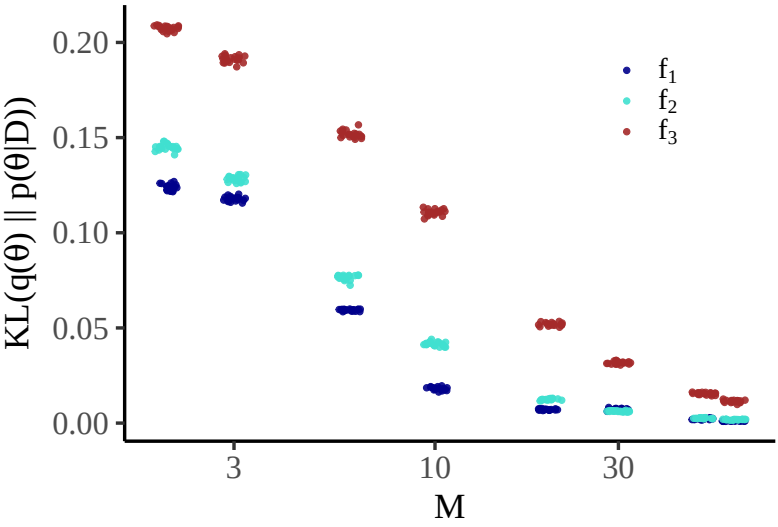

Fig. S3 Effect of the different transformations before the f_{BP} on the performance. In f_1, f_2 $z \sim N(0,1)$ and the transformation are $f_1 = f_{\text{BP}} \circ \sigma \circ l$ and $f_2 = f_{\text{BP}} \circ \sigma$. For f_3, z is sampled from a truncated normal with shape 0.5, scale 0.15, and with bound 10^{-4} to $1 - 10^{-4}$ and transformed via f_{BP}.

training and better fitting with fewer parameters M when additionally prepending an affine transformation function $l = \alpha \cdot z + \beta$ before σ so that the total transformation is $f = f_{\text{BP}} \circ \sigma \circ l$ (see section B.2). We also experimented with an additional affine function after the Bernstein polynomials but did not find a beneficial effect. We use $f = f_{\text{BP}} \circ \sigma \circ l$ in the following, except in B.2 where we studied other possibilities to restrict the input to f_{BP} to $[0, 1]$. Moreover, a distribution F_Z with support $[0, 1]$ can be used. We experimented with a truncated Gaussian.

While a for large number M of the performance for the different methods become similar (see Fig. S3), the transformation $f_1 = f_{\text{BP}} \circ \sigma \circ l$ works better for smaller of M, we thus choose this transformation for all experiments.

B.3 Toy Linear Regression Example

In this example, a posterior is generated, which displays a strong correlation between the results.

The Data Generation Process

The data is generated by the following R-code.

```
set.seed(42)
N = 6L # number data points
```

```
P = 2L #number of predictors
z_dg = rnorm(N, mean=0, sd=1)
x_dg = scale(rnorm(N, mean = 1.42*z_dg, sd = 0.1),
             scale = FALSE)
y_r = rnorm(N, mean = 0.42*x_dg - 1.1 * z_dg, sd=0.42)
```

The Actual Data

```
>    x_r
              [,1]         [,2]
[1,]  1.3709584  1.48475156
[2,] -0.5646982 -1.42449894
[3,]  0.3631284  0.10432308
[4,]  0.6328626  0.27923186
[5,]  0.4042683  0.09138635
[6,] -0.1061245 -0.53519391
>    y_r
[1] -1.46778013 -0.09421285 -0.41162052
-0.31177232 -0.52569912 -1.22375575
```

Details of the Model

The model is a linear regression with Gaussian noise, defined by the likelihood $y \sim \mathcal{N}(xw + b, \sigma)$. The priors are specified as $b \sim \mathcal{N}(0, 10)$, $w \sim \mathcal{N}(0, 10)$, and $\sigma \sim$ Lognormal$(0.5, 1)$.

Details of the Training Procedure

Figure S7 presents the scaled loss curves for the first run of the remaining experiments. Plotted are the epochs versus the negative ELBO, where the curve is normalized by first subtracting the minimum value observed across all epochs and then dividing by the absolute value of this minimum. For the remaining 4 runs, similar behavior is observed. To be comparable with existing literature the ELBO is estimated with $S = 10$ MC samples introducing some level of noise into the data.

B.3.1 Additional Results

Figure S4, shows the result for MF-Gaussian approximation. As is visible, the mean-field approach cannot capture the multivariate correlation of the posterior and strongly focuses on the mode of the posterior.

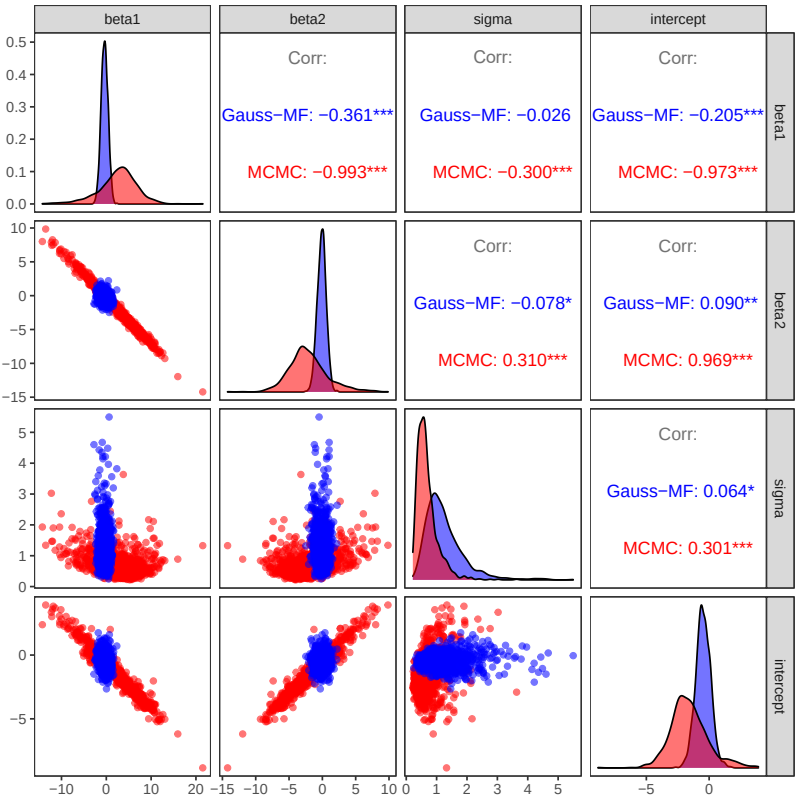

Fig. S4 Mean-Field Gaussian results for the linear regression posteriors.

B.4 Diamond

The diamond data set is modeled with a linear model. To be in accordance with [37], we choose standard normal priors, except for σ and the intercept, for which we choose Student's t-distributions as priors. Specifically, the intercept is assigned a Student's t-distribution with 3 degrees of freedom, a location parameter of 8, and a scale parameter of 10, while σ is modeled using a half-Student's t-distribution with 3 degrees of freedom, a location of 0, and a scale parameter of 10.

B.4.1 Additional Results

In Fig. S5, we show samples from the marginals of the MCMC solution and the BF-VI approximation.

B.5 8School

The 8-Schools example comes in NCP and CP parameterizations. See Table 1 for the model definition, including prios.

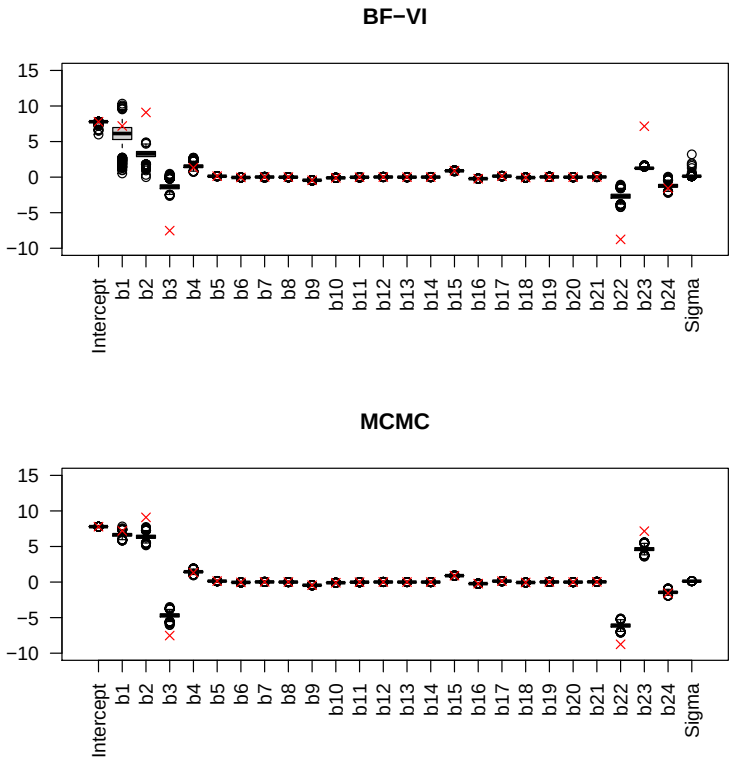

Fig. S5 Comparison of samples from the marginals of MCMC samples and samples from the approximative posterior for the diamond data set. Here, the MCMC solution is very narrow, and the BF-VI (as other approximations) fails to converge to the narrow posterior. The red crosses indicate the maximum likelihood solution.

B.5.1 Additional Results

In Fig. S6, we show samples from the marginals of the 8-dimensional parameter θ (right side, centered parameterization) and $\tilde{\theta}$ (left side, non-centered) parameterization. For the MCMC samples, the easier centered parameterization yielding $\tilde{\theta}$ is used for both cases, and θ is calculated from it. We see an underestimation of the true posterior.

B.6 NN-Based Non-Linear Regression Example

The Actual Data

We have $N = 9$ samples with
$x = -5.412390, -4.142973, -5.100401, -4.588446, -2.057037, -2.003591, -3.740448, -5.344997, 4.506781$ **and**
$y = 0.9731220, 0.9664410, 1.2311585, 0.5193988, -1.2059958, -0.9434611, 0.8041748, 0.8299642, -1.3704962$**.**

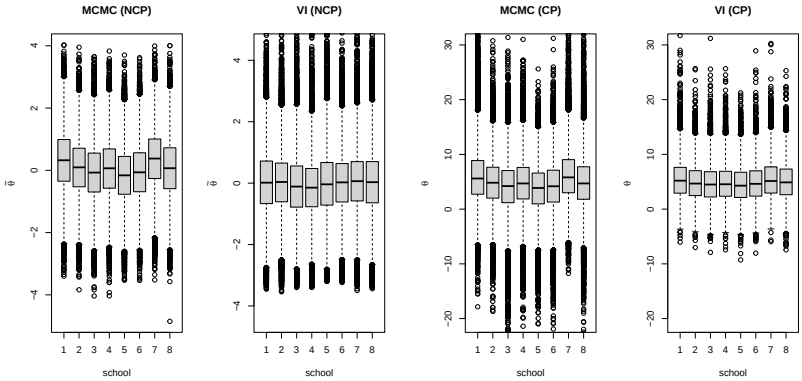

Fig. S6 Marginal samples from MCMC and BF-VI for the 8 schools examples. For the NCP-parametrization $\tilde{\theta}_i$ $(i = 1, \ldots, 8)$ is shown on left two plots and θ right two plots (CP-Parametrization).

Details of the Model

The model is a simple NN with one hidden layer comprising 3 neurons and one neuron in the output layer, giving the conditional mean. All weights have a $N(0, 1)$ prior.

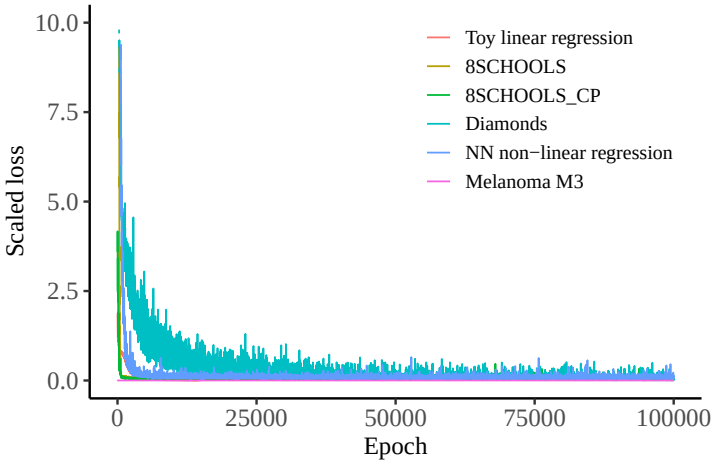

Fig. S7 Loss curve for the different experiments. The loss is scaled by subtracting the minimum loss and dividing by the absolute value of the minimum loss.