A. Other results not included in the paper

In fig. 3 we report some of the runs that we did not include in the main text for lack of space. The figure reports plots on the error vs. time for the same regression cases considered in the main text but with an isotropic kernel, and results on the concrete dataset with isotropic and ARD kernels.

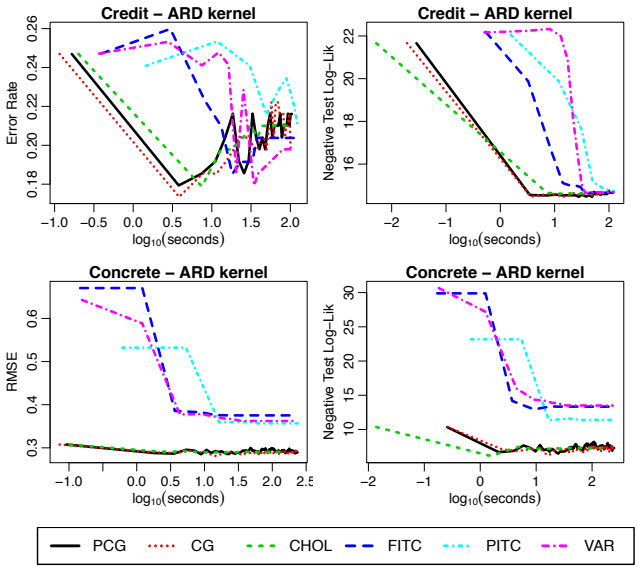

Figure 3. RMSE and negative log of the likelihood on $\sqrt{n}$ held out test data over time. GP models employ the ARD kernel in eq. 1. GP classification: Credi dataset ($n = 1000, d = 24$). GP regression: Concrete dataset ($n = 1029, d = 8$). Curves are averaged over multiple repetitions.

B. Gaussian Processes with non-Gaussian likelihood functions

In this section we report the derivations of the quantities needed to compute an unbiased estimate of the log-marginal likelihood given by the Laplace approximation for GP models with non-Gaussian likelihood functions. Throughout this section, we assume a factorizing likelihood

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i).$$

and we specialize the equations to the probit likelihood

$$p(y_i \mid f_i) = \Phi(y_i f_i). \tag{3}$$

where Φ denotes the cumulative function of the Gaussian density. The latent variables $\mathbf{f}$ are given a zero mean GP prior $\mathbf{f} \sim \mathcal{N}(\mathbf{f}|\mathbf{0}, K)$.

For a given value of the hyperparameters $\boldsymbol{\theta}$, define

$$\Psi(\mathbf{f}) = \log[p(\mathbf{y} \mid \mathbf{f})] + \log[p(\mathbf{f} \mid \boldsymbol{\theta})] + \text{const}. \tag{4}$$

as the logarithm of the posterior density over $\mathbf{f}$. Performing a Laplace approximation amounts in defining a Gaussian $q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f} \mid \hat{\mathbf{f}}, \hat{\Sigma})$, such that

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} \Psi(\mathbf{f}) \qquad \text{and} \qquad \hat{\Sigma}^{-1} = -\nabla_{\mathbf{f}}\nabla_{\mathbf{f}}\Psi(\hat{\mathbf{f}}). \tag{5}$$

As it is not possible to directly solve the maximization problem in equation 5, an iterative procedure based on the following Newton-Raphson formula is usually employed,

$$\mathbf{f}^{\text{new}} = \mathbf{f} - (\nabla_{\mathbf{f}}\nabla_{\mathbf{f}}\Psi(\mathbf{f}))^{-1}\nabla_{\mathbf{f}}\Psi(\mathbf{f}), \tag{6}$$

starting from some initial $\mathbf{f}$ until convergence. The gradient and the Hessian of the log of the target density are

$$\nabla_{\mathbf{f}}\Psi(\mathbf{f}) = \nabla_{\mathbf{f}} \log[p(\mathbf{y} \mid \mathbf{f})] - K^{-1}\mathbf{f} \quad \text{and} \tag{7}$$

$$\nabla_{\mathbf{f}}\nabla_{\mathbf{f}}\Psi(\mathbf{f}) = \nabla_{\mathbf{f}}\nabla_{\mathbf{f}} \log[p(\mathbf{y} \mid \mathbf{f})] - K^{-1} = -W - K^{-1}, \tag{8}$$

where we have defined $W = -\nabla_{\mathbf{f}}\nabla_{\mathbf{f}} \log[p(\mathbf{y} \mid \mathbf{f})]$, which is diagonal because the likelihood factorizes over observations. Note that if $\log[p(\mathbf{y} \mid \mathbf{f})]$ is concave, such as in probit classification, $\Psi(\mathbf{f})$ has a unique maximum.

Standard manipulations lead to

$$\mathbf{f}^{\text{new}} = (K^{-1} + W)^{-1}(W\mathbf{f} + \nabla_{\mathbf{f}} \log[p(\mathbf{y} \mid \mathbf{f})]).$$

We can rewrite the inverse of the negative Hessian using the matrix inversion lemma:

$$\left(K^{-1} + W\right)^{-1} = K - KW^{\frac{1}{2}}B^{-1}W^{\frac{1}{2}}K,$$

where

$$B = I + W^{\frac{1}{2}}KW^{\frac{1}{2}}.$$

This means that each iteration becomes:

$$\mathbf{f}^{\text{new}} = (K - KW^{\frac{1}{2}}B^{-1}W^{\frac{1}{2}}K)(W\mathbf{f} + \nabla_{\mathbf{f}} \log[p(\mathbf{y} \mid \mathbf{f})]).$$

We can define $\mathbf{b} = (W\mathbf{f} + \nabla_{\mathbf{f}} \log[p(\mathbf{y} \mid \mathbf{f})])$ and rewrite this expression as:

$$\mathbf{f}^{\text{new}} = K(\mathbf{b} - W^{\frac{1}{2}}B^{-1}W^{\frac{1}{2}}K\mathbf{b}).$$

From this, we see that at convergence

$$\mathbf{a} = K^{-1}\hat{\mathbf{f}} = (\mathbf{b} - W^{\frac{1}{2}}B^{-1}W^{\frac{1}{2}}K\mathbf{b}).$$

As we will see later, the definition of $\mathbf{a}$ is useful for the calculation of the gradient and for predictions.

Proceeding with the calculations from right to left we see that in order to complete a Newton-Raphson iteration the expensive operations are: (i) carry out one matrix-vector multiplication $K\mathbf{b}$, (ii) solve a linear system involving the

Algorithm 2 Laplace approximation for GPs

Input: data X, labels $\mathbf{y}$, likelihood function $p(\mathbf{y} \mid \mathbf{f})$
$\mathbf{f} = \mathbf{0}$
repeat
  Compute $\mathrm{diag}(W)$, $\mathbf{b}$, $W^{\frac{1}{2}}K\mathbf{b}$
  $\mathrm{solve}(B, W^{\frac{1}{2}}K\mathbf{b})$
  Compute $\mathbf{a}$, $K\mathbf{a}$
  Compute $\mathbf{f}^{\mathrm{new}}$
until convergence
return $\hat{\mathbf{f}}$, $\mathbf{a}$

Algorithm 3 Stochastic gradients for GPs

Input: data X, labels $\mathbf{y}$, $\hat{\mathbf{f}}$, $\mathbf{a}$
$\mathrm{solve}(B, \mathbf{r}^{(i)})$ for $i = 1, \ldots, N_{\mathbf{r}}$
Compute first term of $\tilde{g}_i$
Compute second term of $\tilde{g}_i$
$\mathrm{solve}(B, W^{\frac{1}{2}}K\mathbf{r}^{(i)})$ for $i = 1, \ldots, N_{\mathbf{r}}$
Compute $\tilde{\mathbf{u}}$
$\mathrm{solve}(B, W^{\frac{1}{2}}\frac{\partial K}{\partial \theta_i}\nabla_{\hat{\mathbf{f}}} \log[p(\mathbf{y} \mid \hat{\mathbf{f}})])$
Compute third term of $\tilde{g}_i$
return $\tilde{\mathbf{g}}$

B matrix, and (iii) carry out one matrix-vector multiplication involving K and the vector in the parenthesis. Calculating $\mathbf{b}$ and performing any multiplications of $W^{\frac{1}{2}}$ with vectors cost $\mathcal{O}(n)$.

All these operations can be carried out without the need to store K or any other $n \times n$ matrices. The linear system in (ii) can be solved using the CG algorithm that involves repeatedly multiplying B (and therefore K) with vectors.

B.1. Stochastic gradients

The Laplace approximation yields an approximate log-marginal likelihood in the following form:

$$\log[\hat{p}(\mathbf{y} \mid \boldsymbol{\theta}, X)] = -\frac{1}{2}\log|B| - \frac{1}{2}\hat{\mathbf{f}}^\top K^{-1}\hat{\mathbf{f}} + \log[p(\mathbf{y} \mid \hat{\mathbf{f}})] \tag{9}$$

Handy relationships that we will be using in the remainder of this section are:

$$\log|B| = \log|I + W^{\frac{1}{2}}KW^{\frac{1}{2}}| = \log|I + KW|;$$
$$(I + KW)^{-1} = W^{-\frac{1}{2}}B^{-1}W^{\frac{1}{2}}.$$

The gradient of the log-marginal likelihood with respect to the kernel parameters $\boldsymbol{\theta}$ requires differentiating the terms that explicitly depend on $\boldsymbol{\theta}$ and those that implicitly depend on it because a change in the parameters reflects in a change in $\hat{\mathbf{f}}$. Denoting by g_i the ith component of the gradient of $\frac{\partial \log[\hat{p}(\mathbf{y}|\boldsymbol{\theta})]}{\partial \theta_i}$, we obtain

$$\begin{aligned} g_i &= -\frac{1}{2}\mathrm{Tr}\left(B^{-1}\frac{\partial B}{\partial \theta_i}\right) \\ &\quad + \frac{1}{2}\hat{\mathbf{f}}^\top K^{-1}\frac{\partial K}{\partial \theta_i}K^{-1}\hat{\mathbf{f}} \\ &\quad + \left[\nabla_{\hat{\mathbf{f}}} \log[\hat{p}(\mathbf{y}|\boldsymbol{\theta})]\right]^\top \frac{\partial \hat{\mathbf{f}}}{\partial \theta_i} \end{aligned} \tag{10}$$

The trace term cannot be computed exactly for large n so we propose a stochastic estimate:

$$-\frac{1}{2}\left[\mathrm{Tr}\left(\widetilde{B^{-1}\frac{\partial B}{\partial \theta_i}}\right)\right] = -\frac{1}{2N_{\mathbf{r}}}\sum_{i=1}^{N_{\mathbf{r}}}(\mathbf{r}^{(i)})^\top B^{-1}\frac{\partial B}{\partial \theta_i}\mathbf{r}^{(i)}.$$

By noticing that the derivative of B is $W^{\frac{1}{2}}\frac{\partial K}{\partial \theta_i}W^{\frac{1}{2}}$, this simplifies to

$$-\frac{1}{2N_{\mathbf{r}}}\sum_{i=1}^{N_{\mathbf{r}}}(\mathbf{r}^{(i)})^\top B^{-1}W^{\frac{1}{2}}\frac{\partial K}{\partial \theta_i}W^{\frac{1}{2}}\mathbf{r}^{(i)},$$

so we need to solve $N_{\mathbf{r}}$ linear systems involving B.

The second term contains the linear system $K^{-1}\hat{\mathbf{f}}$ that we already have from the Laplace approximation and is $\mathbf{a}$.

The third term is slightly more involved and will be dealt with in the next sub-section.

B.1.1. Implicit derivatives

The last (implicit) term in the last equation can be simplified by noticing that:

$$\log[\hat{p}(\mathbf{y} \mid \boldsymbol{\theta})] = \Psi(\hat{\mathbf{f}}) - \frac{1}{2}\log|B|$$

and that the derivative of the first term wrt $\hat{\mathbf{f}}$ is zero because $\hat{\mathbf{f}}$ maximizes $\Psi(\hat{\mathbf{f}})$. Therefore:

$$\left[\nabla_{\hat{\mathbf{f}}} \log[\hat{p}(\mathbf{y} \mid \boldsymbol{\theta})]\right]^\top \frac{\partial \hat{\mathbf{f}}}{\partial \theta_i} = -\frac{1}{2}\left[\nabla_{\hat{\mathbf{f}}} \log|B|\right]^\top \frac{\partial \hat{\mathbf{f}}}{\partial \theta_i}$$

The components of $\left[\nabla_{\hat{\mathbf{f}}} \log|B|\right]$ can be obtained by considering the identity $\log|B| = \log|I + KW|$, so differentiating $\log|B|$ wrt the components of $\hat{\mathbf{f}}$ becomes:

$$\frac{\partial \log|I + KW|}{\partial(\hat{\mathbf{f}})_j} = \mathrm{Tr}\left((I + KW)^{-1}K\frac{\partial W}{\partial(\hat{\mathbf{f}})_j}\right)$$

We can rewrite this by gathering K inside the inverse and, due to the inversion of the matrix product, K cancels out:

$$\frac{\partial \log|I + KW|}{\partial(\hat{\mathbf{f}})_j} = \mathrm{Tr}\left((K^{-1} + W)^{-1}\frac{\partial W}{\partial(\hat{\mathbf{f}})_j}\right)$$

We notice here that the resulting trace contains the inverse of the same matrix needed in the iterations of the Laplace approximation and that the matrix $\frac{\partial W}{\partial(\hat{\mathbf{f}})_j}$ is zero everywhere

except in the jth diagonal element where it attains the value:

$$\frac{\partial W}{\partial(\hat{\mathbf{f}})_j} = \frac{\partial^3 \log[p(\mathbf{y} \mid \hat{\mathbf{f}})]}{\partial(\hat{\mathbf{f}})_j^3}$$

For this reason, it would be possible to simplify the trace term as the product between the jth diagonal element of $(K^{-1}+W)^{-1}$ and $\frac{\partial^3 \log[p(\mathbf{y}|\hat{\mathbf{f}})]}{\partial(\hat{\mathbf{f}})_j^3}$. Bearing in mind that we need n of these quantities, we could define

$$D = \text{diag}\left[\text{diag}\left[(K^{-1}+W)^{-1}\right]\right]$$

$$(\mathbf{d})_j = \frac{\partial^3 \log[p(\mathbf{y} \mid \hat{\mathbf{f}})]}{\partial(\hat{\mathbf{f}})_j^3}$$

and rewrite

$$-\frac{1}{2}\left[\nabla_{\hat{\mathbf{f}}} \log|B|\right] = -\frac{1}{2} D\mathbf{d}$$

which is the standard way to proceed when computing the gradient of the approximate log-marginal likelihood using the Laplace approximation (Rasmussen & Williams, 2006). However, this would be difficult to compute exactly for large n, as this would require inverting $K^{-1}+W$ first and then compute its diagonal. Using the matrix inversion lemma would not simplify things as there would still be an inverse of B to compute explicitly. We therefore aim for a stochastic estimate of this term starting from:

$$\begin{aligned}\frac{\partial \log|I+KW|}{\partial(\hat{\mathbf{f}})_j} &= \text{Tr}\left((K^{-1}+W)^{-1}\frac{\partial W}{\partial(\hat{\mathbf{f}})_j}\right) \\ &= \text{Tr}\left((K^{-1}+W)^{-1}\frac{\partial W}{\partial(\hat{\mathbf{f}})_j}\text{E}[\mathbf{r}\mathbf{r}^\top]\right)\end{aligned} \tag{11}$$

where we have introduced the $\mathbf{r}$ vectors with the property $\text{E}[\mathbf{r}\mathbf{r}^\top] = I$. So an unbiased estimate of the trace for each component of $\hat{\mathbf{f}}$ is:

$$\begin{aligned}(\tilde{\mathbf{u}})_j &= \widetilde{\left[\frac{\partial \log|I+KW|}{\partial(\hat{\mathbf{f}})_j}\right]} \\ &= \frac{1}{N_{\mathbf{r}}}\sum_{i=1}^{N_{\mathbf{r}}}(\mathbf{r}^{(i)})^\top(K^{-1}+W)^{-1}\frac{\partial W}{\partial(\hat{\mathbf{f}})_j}\mathbf{r}^{(i)}\end{aligned} \tag{12}$$

which requires solving $N_{\mathbf{r}}$ linear systems involving the B matrix:

$$(K^{-1}+W)^{-1}\mathbf{r}^{(i)} = K(\mathbf{r}^{(i)} - W^{\frac{1}{2}}B^{-1}W^{\frac{1}{2}}K\mathbf{r}^{(i)})$$

The derivative of $\hat{\mathbf{f}}$ wrt θ_i can be obtained by differentiating the expression $\hat{\mathbf{f}} = K\nabla_{\hat{\mathbf{f}}}\log[p(\mathbf{y} \mid \hat{\mathbf{f}})]$:

$$\frac{\partial \hat{\mathbf{f}}}{\partial \theta_i} = \frac{\partial K}{\partial \theta_i}\nabla_{\hat{\mathbf{f}}}\log[p(\mathbf{y} \mid \hat{\mathbf{f}})] + K\nabla_{\hat{\mathbf{f}}}\nabla_{\hat{\mathbf{f}}}\log[p(\mathbf{y} \mid \hat{\mathbf{f}})]\frac{\partial \hat{\mathbf{f}}}{\partial \theta_i}$$

Algorithm 4 Prediction for GPs with Laplace approximation without Cholesky decompositions

1: **Input:** data X, labels $\mathbf{y}$, test input $\mathbf{x}_*$, $\hat{\mathbf{f}}$, $\mathbf{a}$
2: Compute μ_*
3: solve$(B, W^{\frac{1}{2}}\mathbf{k}_*)$
4: Compute s_*^2, $\Phi\left(\frac{m_*}{\sqrt{1+s_*^2}}\right)$
5: **return** $\Phi\left(\frac{m_*}{\sqrt{1+s_*^2}}\right)$

Given that $\nabla_{\hat{\mathbf{f}}}\nabla_{\hat{\mathbf{f}}}\log[p(\mathbf{y} \mid \hat{\mathbf{f}})] = -W$ we can rewrite:

$$(I+KW)\frac{\partial \hat{\mathbf{f}}}{\partial \theta_i} = \frac{\partial K}{\partial \theta_i}\nabla_{\hat{\mathbf{f}}}\log[p(\mathbf{y} \mid \hat{\mathbf{f}})]$$

which yields:

$$\frac{\partial \hat{\mathbf{f}}}{\partial \theta_i} = (I+KW)^{-1}\frac{\partial K}{\partial \theta_i}\nabla_{\hat{\mathbf{f}}}\log[p(\mathbf{y} \mid \hat{\mathbf{f}})]$$

So an unbiased estimate of the implicit term in the gradient of the approximate log-marginal likelihood becomes:

$$-\frac{1}{2}\tilde{\mathbf{u}}^\top(I+KW)^{-1}\frac{\partial K}{\partial \theta_i}\nabla_{\hat{\mathbf{f}}}\log[p(\mathbf{y} \mid \hat{\mathbf{f}})]$$

Rewriting the inverse in terms of B yields:

$$-\frac{1}{2}\tilde{\mathbf{u}}^\top W^{-\frac{1}{2}}B^{-1}W^{\frac{1}{2}}\frac{\partial K}{\partial \theta_i}\nabla_{\hat{\mathbf{f}}}\log[p(\mathbf{y} \mid \hat{\mathbf{f}})]$$

Putting everything together, the components of the stochastic gradient are:

$$\begin{aligned}\tilde{g}_i &= -\frac{1}{2N_{\mathbf{r}}}\sum_{i=1}^{N_{\mathbf{r}}}(\mathbf{r}^{(i)})^\top B^{-1}W^{\frac{1}{2}}\frac{\partial K}{\partial \theta_i}W^{\frac{1}{2}}\mathbf{r}^{(i)} \\ &\quad +\frac{1}{2}\mathbf{a}^\top\frac{\partial K}{\partial \theta_i}\mathbf{a} \\ &\quad -\frac{1}{2}\tilde{\mathbf{u}}^\top W^{-\frac{1}{2}}B^{-1}W^{\frac{1}{2}}\frac{\partial K}{\partial \theta_i}\nabla_{\hat{\mathbf{f}}}\log[p(\mathbf{y}|\hat{\mathbf{f}})]\end{aligned} \tag{13}$$

B.2. Predictions

To obtain an approximate predictive distribution, conditioned on a value of the hyperparameters $\boldsymbol{\theta}$, we can compute:

$$p(y_* \mid \mathbf{y}, \boldsymbol{\theta}) = \int p(y_* \mid f_*)p(f_* \mid \mathbf{f}, \boldsymbol{\theta})q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta})df_*d\mathbf{f}. \tag{14}$$

Given the properties of multivariate normal variables, f_* is distributed as $\mathcal{N}(f_* \mid \mu_*, \beta_*^2)$ with $\mu_* = \mathbf{k}_*^\top K^{-1}\mathbf{f}$ and $\beta_*^2 = k_{**} - \mathbf{k}_*^\top K^{-1}\mathbf{k}_*$. Approximating $p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta})$ with

a Gaussian $q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f} \mid \boldsymbol{\mu}_q, \Sigma_q)$ makes it possible to analytically perform integration with respect to $\mathbf{f}$ in eq. 14. In particular, the integration with respect to $\mathbf{f}$ yields $\mathcal{N}(f_* \mid m_*, s_*^2)$ with

$$m_* = \mathbf{k}_*^\top K^{-1} \hat{\mathbf{f}}$$

and

$$s_*^2 = k_{**} - \mathbf{k}_*^\top (K + W^{-1})^{-1} \mathbf{k}_*$$

These quantities can be rewritten as:

$$m_* = \mathbf{k}_*^\top \mathbf{a}$$

and

$$s_*^2 = k_{**} - \mathbf{k}_*^\top W^{\frac{1}{2}} B^{-1} W^{\frac{1}{2}} \mathbf{k}_*$$

This shows that the mean is cheap to compute, whereas the variance requires solving another linear system involving B for each test point.

The univariate integration with respect to f_* follows exactly in the case of a probit likelihood, as it is a convolution of a Gaussian and a cumulative Gaussian

$$\int p(y_* \mid f_*) \mathcal{N}(f_* \mid m_*, s_*^2) df_* = \Phi\left(\frac{m_*}{\sqrt{1 + s_*^2}}\right). \tag{15}$$

B.3. Low rank preconditioning

When a low rank approximation of the matrix K is available, say $\hat{K} = \Phi\Phi^\top$, the inverse of the preconditioner can be rewritten as:

$$(I + W^{\frac{1}{2}} \hat{K} W^{\frac{1}{2}})^{-1} = (I + W^{\frac{1}{2}} \Phi\Phi^\top W^{\frac{1}{2}})^{-1}$$

By using the matrix inversion lemma we obtain:

$$(I + W^{\frac{1}{2}} \Phi\Phi^\top W^{\frac{1}{2}})^{-1} = I - W^{\frac{1}{2}} \Phi (I + \Phi^\top W \Phi)^{-1} \Phi^\top W^{\frac{1}{2}}$$

Similarly to the GP regression case, the application of this preconditioner is in $\mathcal{O}(m^3)$, where m is the rank of Φ.