## Supplemental Content

## A Bayesian optimization

Bayesian optimization methods [10, 20] try to minimize, over some domain $\mathcal{X} \subset \mathbb{R}^d$, a function $f : \mathcal{X} \to \mathbb{R}$ while sampling $f$ as little as possible (this is what is meant by "sample efficient"). The optimization is generally initialized with $p$ evaluations by sampling with low discrepancy sequences [2] such as latin hypercube sampling.

After the initialization, the sequential component of the optimization begins. At iteration $t$, all previously observed data $\mathbf{y} = \mathbf{y}_{1:t}$ at points $\mathbf{X} = \mathbf{X}_{1:t}$ is used to construct a probabilistic surrogate model $s_{\mathbf{y},\mathbf{X}}$. The next location $\mathbf{x}_{t+1}$ is determined by maximizing a chosen *acquisition function* which measures the benefit or utility associated with evaluating a proposed $\mathbf{x} \in \mathcal{X}$. In this article, we restrict our focus to only considering expected improvement [11],

$$EI(\mathbf{x}) = \mathbb{E}_{p(y|s_{\mathbf{y},\mathbf{X}}(\mathbf{x}))} \left(\max(\mathbf{y}_{1:t}) - y\right)_+ . \tag{Eq. A.1}$$

Most frequently, the probabilistic surrogate model takes the form of a Gaussian Process (GP), although other alternatives have been presented, such as random forests [6], kernel density estimators [1] or Bayesian neural networks [22, 23]. For the remainder of the paper we only consider a GP with zero mean and covariance $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the surrogate model.

**Observation model** We choose to build our GP models on the belief that data has been observed in the presence of homoscedastic noise $Y_{\mathbf{x}} = f(\mathbf{x}) + \epsilon$, with *Gaussian likelihood* $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$, resulting in a GP posterior model. We can also rewrite the likelihood as $Y_{\mathbf{x}}|f \sim \mathcal{N}(f(\mathbf{x}), \sigma_n^2)$. However, as we discuss in Appendix B, this model is not robust to outliers and we will replace the Gaussian likelihood for a more suitable distribution, that is, the Student-$t$.

In this setting, after $t$ observations (as explained in Appendix A), the GP posterior model gives predictions at a query point $\mathbf{x}_q$ which are normally distributed $Y_{\mathbf{x}_q}|\mathbf{X}, \mathbf{y} \sim \mathcal{N}(\mu(\mathbf{x}_q), \sigma^2(\mathbf{x}_q))$, such that

$$\begin{aligned} \mu(\mathbf{x}_q) &= \mathbf{k}(\mathbf{x}_q, \mathbf{X})^T \mathbf{K}^{-1} \mathbf{y}, \\ \sigma^2(\mathbf{x}_q) &= k(\mathbf{x}_q, \mathbf{x}_q) - \mathbf{k}(\mathbf{x}_q, \mathbf{X})^T \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}_q, \mathbf{X}), \end{aligned} \tag{Eq. A.2}$$

where

$$\begin{aligned} \mathbf{k}(\mathbf{x}_q, \mathbf{X}) &= \begin{pmatrix} k(\mathbf{x}_q, \mathbf{x}_1) & \dots & k(\mathbf{x}_q, \mathbf{x}_t) \end{pmatrix}^T, \\ \mathbf{K} &= \begin{pmatrix} \mathbf{k}(\mathbf{x}_1, \mathbf{X}) & \dots & \mathbf{k}(\mathbf{x}_t, \mathbf{X}) \end{pmatrix} + \mathbf{I}\sigma_n^2. \end{aligned} \tag{Eq. A.3}$$

The kernel is chosen to be the Matérn kernel with $\nu = 5/2$, also called $C^4$ Matérn kernel [3],

$$k(\mathbf{x}, \mathbf{x}') = \left(1 + r + r^2/3\right) \mathrm{e}^{-r},$$

where $r = \|\mathbf{x} - \mathbf{x}'\|_\Lambda$ for some positive definite matrix $\Lambda$. The automatic relevance determination kernel which we use here restricts $\Lambda$ to being diagonal. The hyperparameters of $\Lambda$ are estimated by maximum likelihood, although MCMC could be used instead [21].

## B Robust Regression For Gaussian Processes

Standard GP-based Bayesian optimization uses an observation model for noisy data with a Gaussian likelihood, $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ as defined in (Eq. A.3); this allows for closed form inference but, as shown in Figure 1, is very sensitive to outliers. In this section, we review literature on robust regression [17] to draw inspiration on how to alter the standard GP used in Bayesian optimization to create a version robust to outliers.

The key change in creating such a regression model is using a heavy-tailed distribution as the observation likelihood in lieu of the standard Gaussian likelihood; possible options include the Laplace, the hyperbolic secant, or the Student-$t$ likelihoods. All those probability distributions are robust to the presence of outliers, with the Student-$t$ likelihood usually providing the best results

[7, 9]. O'Hagan proved that the Student-$t$ distribution can reject up to $m$ outliers tending to infinity (or negative infinity) provided that there are at least $2m$ observations at all. At the same time, he also showed in [14] that the Gaussian distribution is *nonrobust*, meaning that if an outlier is not rejected, the larger the error present in the outlier, the larger the estimate bias will be.

However, the Student-$t$ likelihood, as well as the alternative distributions mentioned, do not allow closed form inference of the posterior. Therefore, we need to find an approximation that will provide a suitable posterior in the form of a GP or similar.

**Related work:** Vanhatalo et al. suggested to use the Laplace approximation to compute the posterior inference of a GP with Student-$t$ likelihood [27]. The same authors later compared different strategies: MCMC [13], variational inference, and a modification of the expectation propagation (EP) algorithm with double-loop [7]. They showed that their modification of the EP is the most robust estimation method, although it has an increased computational cost. It is important to note that the vanilla version of EP does not converge at all for the Student-$t$ likelihood [16]. In a different approach, Shah et al. [19] had the surprising result that a Student-$t$ process prior with additive noise in the kernel behaves like a Gaussian or Student-$t$ process posterior with a long-tailed likelihood, similar to the Student-$t$ distribution. The surprise arise for the fact that the Student-$t$ process prior is, by definition, robust to input variables $\mathbf{x}$ but not target variables $y$. The advantage of this method is that it is analytical, removing the extra cost of the iterative approximation. However, the actual statistical properties of the method were unclear. This idea was later proved to have the same marginal likelihood as a Student-$t$ process with dependent Student-$t$ noise, giving a probabilistic interpretation of the results [26]. This dependency in the noise might be a strong assumption for certain applications.

Furthermore, the critical parameter controlling robustness of the Student-$t$ distribution it is the degrees of freedom $\nu$, which is recommended to be at least 4 in practice [7, 9]. However, this parameter cannot be tuned independently in the dependent case. Furthermore, in the case of noisy data, learning the noise level is harder in the additive noise model due to the entanglement of the variables. This issue was recently addressed by Tang et al. [25] by using again the Laplace approximation from Vanhatalo et al. [27] to obtain an independent $t$ noise model in a Student-$t$ process.

### B.1 Numerical approximation of Student-$t$ likelihood

First, we will use the Student-$t$ likelihood from Vanhatalo et al. [27]. The Student-$t$ distribution has the form

$$t(y; f, \sigma_0, \nu) = \frac{\Gamma\left(\nu + \frac{1}{2}\right)}{\Gamma(\frac{\nu}{2})\sqrt{(\nu\pi)\sigma_0}} \left[1 + \frac{(y-f)^2}{\nu\sigma_0^2}\right]^{-\nu-\frac{1}{2}},$$

where $f \equiv f(\mathbf{x})$, $\nu$ is the degrees of freedom and $\sigma_0$ is the scale parameter. In the Bayesian context, the Student-$t$ distribution usually arises from a normal distribution with a conjugate hyperprior on the variance variable, such as the inverse-$\chi^2$, the inverse gamma or even the Jeffreys prior [18]. For example, in this case, the model $y|f \sim t(y; f, \sigma_0, \nu)$ is equivalent to the original Gaussian likelihood with an hyperprior on the noise term $\sigma_n$. That is:

$$\begin{aligned} y|f, \sigma_n &\sim \mathcal{N}(f, \sigma_n^2) \\ \sigma_n^2|\nu, \sigma_0^2 &\sim \chi^{-2}(\nu, \sigma_0^2) \end{aligned} \qquad \text{(Eq. B.1)}$$

Jylanki et at. [7] present different approximation methods for which we implemented the Laplace method (the simplest and most extended method) [27]. These works were mostly intended for regression applications where large amounts of data are available at once. In contrast, Bayesian optimization seeks to minimize the number of data points, often resulting in less data than most regression applications. Furthermore, observations arrive sequentially. In this context, we found the Laplace approximation to be reliable and numerically stable, because the lack of data resulted in a regularization effect. We also implemented the modified double-loop EP algorithm from [7], but preliminary results resulted in poor performance with many iterations converging to the wrong solution or not converging at all. We conjecture that this effect is because of the limited data available, and further research is required. For brevity, we present results in Section 3 using only the Laplace approximation.

We are interested in computing the predictive posterior from equation (Eq. A.2) with the new likelihood function. The Laplace approximation for the conditional posterior of the latent function,

which we write as $p(f|\mathbf{y}, \mathbf{X}, \Lambda, \sigma_0^2, \nu)$, is constructed from the second order Taylor expansion of log posterior around the mode $\hat{f}$, which results in a Gaussian approximation:

$$p(f|\mathbf{y}, \mathbf{X}, \Lambda, \sigma_0^2, \nu) \approx \mathcal{N}(f|\hat{f}, \Sigma), \tag{Eq. B.2}$$

where $\hat{f} = \arg\max p(f|\mathbf{y}, \mathbf{X}, \Lambda, \sigma_0^2, \nu)$ is the maximum *a posteriori* and $\Sigma^{-1} = \mathbf{K}^{-1} + \mathbf{W}$ the Hessian of the negative log conditional posterior at the mode with, $\mathbf{W} = diag_i \left( \nabla_{f_i} \nabla_{f_i} \log p(y|f_i, \sigma, \nu)|_{f_i = \hat{f}_i} \right)$.

Finally, the new predictive distribution can be computed by marginalization of equation (Eq. B.2). That is:

$$\mu(\mathbf{x}_q) = \mathbf{k}^T \mathbf{K}^{-1} \hat{f},$$
$$\sigma^2(\mathbf{x}_q) = k - \mathbf{k}^T \left( \mathbf{K} + \mathbf{W}^{-1} \right)^{-1} \mathbf{k},$$

where $k = k(\mathbf{x}_q, \mathbf{x}_q)$ and $\mathbf{k} = \mathbf{k}(\mathbf{x}_q, \mathbf{X})$. We refer to Vanhatalo et al. [27] for implementation details.

### B.2 Student-$t$ process with additive kernel noise

For comparison, we also include the Student-$t$ process from Shah et al. [19], which we will the define in terms of the conditional posterior in the form of a multivariate Student-$t$ distribution. This process completely changes the model presented in Appendix A. For brevity we do not include the equations for the predictive distribution, hyperparmenter optimization and expected improvement with the new model. These can be found in the literature [19, 10, 18, 29].

In this case, the Student-$t$ process is generated by placing an inverse gamma prior[1] on the scale parameter of the kernel matrix [10], that is, at the stage we replace the kernel matrix from equation (Eq. A.3) to

$$\mathbf{K} = \sigma_s^2 \left[ (\mathbf{k}(\mathbf{x}_1, \mathbf{X}) \quad \dots \quad \mathbf{k}(\mathbf{x}_t, \mathbf{X})) + \mathbf{I}\sigma_n^2 \right]$$

with $\sigma_s^2 \sim \mathcal{IG}(a, b)$. This is the multivariate generalization of equation (Eq. B.1). Note also how the signal $\sigma_s^2$ and noise $\sigma_n^2$ variances become entangled. As reported by Shah et al. [19], this results are analogous to the method of using the inverse Wishart process as a prior on $\mathbf{K}$. The multivariate Student-$t$ distribution that generate the corresponding process is defined as:

$$t(\mathbf{y}; m, \mathbf{\Sigma}, a, b) = \frac{\Gamma\left(a + \frac{n}{2}\right)}{\Gamma(a)} \frac{1}{\sqrt{(2a\pi)^n \left|b^{-1}\mathbf{K}\right|}} \left[ 1 + \frac{b(\mathbf{y} - m)^T \mathbf{K}^{-1} (\mathbf{y} - m)}{2a} \right]^{-a - \frac{n}{2}}$$

where $m = m(\mathbf{x})$ is the mean function, which is generally assumed to be $m(\mathbf{x}) = 0$ and $a$ and $b$ are the parameters of the inverse gamma. Again, we refer to the literature for implementation details about the posterior inference [19, 10, 18, 29].

## C Outlier Diagnostics in our Filtering

This part of our method is independent of the robust regression model selected before (see Appendix B.1 and Appendix B.2), although for clarity we will assume that we are using the GP with Student-$t$ likelihood from Appendix B.1. Once we have built the robust regression model, we are able to identify the outliers from the rest of the data. As can be seen in Figure 1, the mean function computed with the robust regression (center) is not biased like the nonrobust regression (left). Therefore, we can determine that the outliers are the points in the tail of the predictive distribution. For example, note how the point close to $(-2, 2)$ introduces a large bias in the nonrobust regression. As a result, in the nonrobust regression, the mean prediction is much closer to the point.

[1] Note that the inverse gamma is also equivalent to the scaled inverse $\chi^2$ distribution $\chi_\nu^{-2}(\sigma_0^2) = \mathcal{IG}\left(\frac{\nu}{2}, \frac{\nu\sigma_0^2}{2}\right)$, which will define the Student-$t$ in terms of the degrees of freedom [15] as mentioned in Appendix B.

For that purpose, we compute the upper and lower $\alpha$-percentile of the predicted distribution as a classification threshold, where $\alpha$ is the assumed level of outliers. The selection of this parameter will determine the number of false positives and false negatives. High values of $\alpha$ will classify many points as outliers, reducing the effective sample size for Bayesian optimization. On the other hand, low values of $\alpha$ reduce the robustness of the method by misclassifying actual outliers.

**No permanent rejection** In theory, assuming that a single observation arrives per iteration, only that last observation should be questioned. However, because new data helps improve the model, we found that reclassifying all the points worked better, as new information allows better classification over past observations. Sometimes, points that initially were considered outliers can be found part of the model while, more frequently, points that were initially misclassified as acceptable are properly detected with a better model. For Bayesian optimization, the general assumption is that data points are *expensive* in some sense (such as time or energy), thus no point is permanently deleted or ignored.

**Scheduling diagnostics** Although the Student-$t$ likelihood is able to identify $m$ outliers out of $2m$ points, we have found that in practices it is reasonable to wait for a certain number of iterations before starting classifying data. The motivation is to have a proper regression model with a correct estimate of the hyperparameters. We also found that, because of the sequential nature of Bayesian optimization, if the last point is misclassified as an outlier and removed in the Bayesian optimization, there is a large probability that the will be selected again in the next iteration, which will again might result in a misclassification and so on, wasting valuable resources. Finally, the computational cost of the Student-$t$ likelihood is more expensive than the Gaussian likelihood. Therefore, we propose to use the Student-$t$ likelihood and posterior data filtering after $n_{\text{init}}$ points and, then only once out of each $n_s$ subsequent iterations.

Finally, once the outliers are classified and removed, the optimization is performed with a standard GP computed only with the remaining points, because it produces more stable and fast solutions (see Figure 1). This proved especially true at early stages, when the regression model is noisy and inaccurate, and some large misclassifications might happen. Knowing that there is a limitation on the number of outliers that the Student-$t$ distribution is robust, we are able to detect if there has been a failure in the filtering process by checking if the number of outliers is larger than $m$ for a total of $2m$ points.