Bayesian Quantile and Expectile Optimisation

Abstract

Bayesian optimisation (BO) is widely used to optimise stochastic black box functions. While most BO approaches focus on optimising conditional expectations, many applications require risk-averse strategies and alternative criteria accounting for the distribution tails need to be considered. In this paper, we propose new variational models for Bayesian quantile and expectile regression that are well-suited for heteroscedastic noise settings. Our models consist of two latent Gaussian processes accounting respectively for the conditional quantile (or expectile) and the scale parameter of an asymmetric likelihood functions. Furthermore, we propose two BO strategies based on entropy search and Thompson sampling, that are tailored to such models and that can accommodate large batches of points. Contrary to existing BO approaches for risk-averse optimisation, our strategies can directly optimise for the quantile and expectile, without requiring replicating observations or assuming a parametric form for the noise. As illustrated in the experimental section, the proposed approach clearly outperforms the state of the art in the heteroscedastic, non-Gaussian case.

1 INTRODUCTION

Let $\Psi : \mathcal{X} \times \Omega \to \mathbb{R}$ be an unknown function, where $\mathcal{X} \subset [0,1]^D$ and Ω denotes a probability space representing some uncontrolled variables. For any fixed $x \in \mathcal{X}$, $Y_x = \Psi(x, \cdot)$ is a random variable of distribution $\mathbb{P}_x$. We assume here a classical *black-box optimisation* framework: Ψ is available only through (costly) pointwise evaluations of Y_x. Typical examples may include stochastic simulators in physics or biology (see Skullerud (1968) for simulations of ion motion and Székely Jr and Burrage (2014) for simulations of heterogeneous natural systems), but Ψ can also represent the performance of a machine learning algorithm according to some hyperparameters (see Bergstra et al. (2011) for instance). In the latter case, the randomness can come from the use of minibatching in the training procedure, the choice of a stochastic optimiser or the randomness in the initialisation of the optimiser.

Let $g(x) = \rho(\mathbb{P}_x)$ be the objective function we want to maximise, where ρ is a real-valued functional defined on probability measures. The canonical choice for ρ is the expectation, which is sensible when the exposition to extreme values is not a significant aspect of the decision. However, in a large variety of fields such as agronomy, medicine or finance, decision makers have an incentive to protect themselves against extreme events since they may lead to severe consequences. To take these rare events into account, one should consider alternative choices for ρ that can capture the behaviour of the tails of $\mathbb{P}_x$, such as the quantile (Rostek, 2010), conditional value-at-risk (CVaR, see Rockafellar et al. (2000)) or expectile (Bellini and Di Bernardino, 2017). In this paper we focus our interest on the modelling and optimisation of quantiles and expectiles.

Given an estimate of g based on available data, global optimisation algorithms define a policy that finds a trade-off between exploration and intensification. More precisely, the algorithm has to explore the input space in order to avoid getting trapped in a local optimum, but it also has to concentrate its budget on input regions identified as having a high potential. The latter results in accurate estimates of g in the region of interest and allows the algorithm to return an optimal input value with high precision.

In the context of Bayesian optimisation (BO), such trade-offs have been initially studied by Mockus et al. (1978) and Jones et al. (1998) in a noise-free setting. Their framework has latter been extended to optimisation of the conditional expectation of a stochastic black box (see e.g. Frazier et al. (2009); Srinivas et al. (2009) or Picheny et al. (2013) for a review). Recently, strategies optimising risk measures have been proposed, In particular, Cakmak et al. (2020) proposed new algorithms to optimise for the quantile and CVaR for a slightly different use case, where the space Ω is actually controllable. Browne et al. (2016) and Makarova et al. (2021) proposed algorithms to optimise quantiles and CVaRs, but both rely on intensively repeating observations, which hinders their efficiency in a relatively low budget scenario.

Contributions The contributions of this paper are the following: 1) We propose a new model based on two latent Gaussian Processes (GPs) to estimate quantiles or expectiles that is tailored to heteroscedastic noise. 2) We use Sparse posterior and variational inference to support potentially large datasets. 3) We propose a new Bayesian algorithm suited to optimise conditional quantiles or expectiles in a data efficient manner. Two batch-sequential acquisition strategies are designed to find a good trade-off between exploration and intensification. The ability of our algorithm to optimise quantiles is illustrated on multiple test problems.

2 BAYESIAN METAMODELS OF RISK MEASURES

For a given input point x, the quantile of order $\tau \in (0,1)$ of Y_x can be defined as

$$q_\tau(x) = \underset{q\in\mathbb{R}}{\arg\min}\, \mathbb{E}\big[l_\tau(Y_x - q)\big], \tag{1}$$

where l_τ is the pinball loss (Koenker and Bassett Jr, 1978)

$$l_\tau(\xi) = (\tau - \mathbb{1}_{(\xi<0)})\xi, \quad \xi \in \mathbb{R}. \tag{2}$$

Similarly, Newey and Powell (1987) introduced the expectile as the minimiser of an asymmetric quadratic loss:

$$e_\tau(x) = \underset{q\in\mathbb{R}}{\arg\min}\, \mathbb{E}\big[l^e_\tau(Y_x - q)\big], \tag{3}$$

$$l^e_\tau(\xi) = |\tau - \mathbb{1}_{(\xi<0)}|\xi^2, \quad \xi \in \mathbb{R}. \tag{4}$$

We detail in the next section how these losses can be used to get an estimate of the objective function $g(x)$ using a dataset $\mathcal{D}_n = \big((x_1,y_1)\cdots,(x_n,y_n)\big) = (\mathcal{X}_n,\mathcal{Y}_n)$ that does not necessarily require replicates of observations at the same input location.

2.1 QUANTILE AND EXPECTILE METAMODEL

Different metamodels have been proposed to estimate a quantile function, such as artificial neural networks (Cannon, 2011), random forest (Meinshausen, 2006) or nonparametric estimation in reproducing kernel Hilbert spaces (Takeuchi et al., 2006). While the literature on expectile regression is less extended, neural network (Jiang et al., 2017) or SVM-like approaches (Farooq and Steinwart, 2017) have been developed as well. All the approaches cited above defined an estimator of g as the function that minimises (optionally with a regularisation term)

$$\mathcal{R}_e[g] = \frac{1}{n}\sum_{i=1}^{n} l\big(y_i - g(x_i)\big), \tag{5}$$

with $l = l_\tau$ for the quantile estimation and $l = l^e_\tau$ for the expectile. This framework makes sense because asymptotically minimising (5) is equivalent to minimising (1) or (3).

These approaches however share a common drawback: they do not capture the uncertainty associated with each prediction. This is a significant problem in our setting since quantifying this uncertainty is of paramount importance to define the exploration/intensification trade-off. This limitation can be overcome by using a probabilistic model such as

$$y = g(x) + \epsilon(x),$$

where g is either an unknown parametric function (Yu and Moyeed, 2001) or a Gaussian process (Boukouvalas et al., 2012; Abeywardana and Ramos, 2015), and where the distribution of ϵ depends on the quantity to be estimated. For modelling a quantile, ϵ should follow an asymmetric Laplace distribution:

$$p_\epsilon(e) = \frac{\tau(1-\tau)}{\sigma}\exp\left(-\frac{l_\tau(e)}{\sigma}\right).$$

For approximating an expectile, one can use the asymmetric Gaussian distribution:

$$p_\epsilon(e) = C(\tau,\sigma)\exp\left(-\frac{l^e_\tau(e))}{2\sigma^2}\right), \tag{6}$$

with $C(\tau,\sigma) = \dfrac{\sqrt{2\tau(1-\tau)}}{\sigma\sqrt{\pi}(\sqrt{\tau}+\sqrt{1-\tau})}$.

In both cases, the associated likelihood is given by

$$p(\mathcal{Y}_n|g) = \prod_{i=1}^{n} p_\epsilon(y_i - g(x_i)). \tag{7}$$

Although the Bayesian quantile model presented above is well known (Yu and Moyeed, 2001; Boukouvalas et al., 2012; Abeywardana and Ramos, 2015), the Bayesian expectile model we just introduced is new to the best of our knowledge. It is worth noting that the non-conjugacy between the prior on g and the likelihood functions implies that the posterior distribution of g given the data is not available in closed form. To overcome this, Boukouvalas et al. (2012) use Expectation propagation whereas Abeywardana and Ramos (2015) favours variational inference. The latter appears to be one of the most competitive approaches on the benchmark presented in Torossian et al. (2019) so we will embrace the variational inference framework in the remaining of the paper.

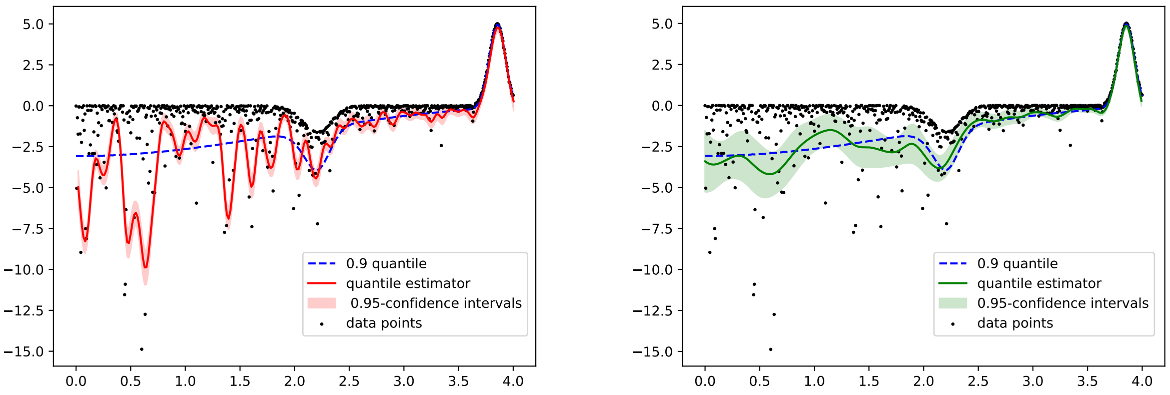

Figure 1: GP quantile model from Abeywardana and Ramos (2015) (left) and ours (right) on data with high heteroscedasticy. The left model cannot compromise between very small observation variances around $x = 4$ and very large variances ($x \leq 2$), largely overfits on half of the domain and returns overconfident confidence intervals. In contrast, our model captures both the low and high variance regions, while returning well-calibrated confidence intervals.

One limitation of the aforementioned methods is that they can result in overconfident predictions in heteroscedastic settings, as illustrated in Figure 1. The main reason is that they only use a single parameter σ to capture the spread for the likelihood function, which amounts to considering that the noise amplitude does not change over the input space. We believe this can be a severe limitation in the context of quantile optimisation since the fluctuation of the quantile value over the input space is likely to be dictated by the noise distribution itself not being stationary.

To overcome this issue, we propose to build quantile and expectile models where the spread of the asymmetric Laplace and Gaussian likelihoods varies across the input space. For both distributions, this can be achieved by redefining σ in equations 7 and 6 as a function of the input parameters. Intuitively, a small value of $\sigma(x)$ means that there is a high penalty for having an estimate of $g(x)$ that is far away from the data, whereas a large value of $\sigma(x)$ means that this penalty is limited and thus leads to more regularity in the model predictions. In practice, we choose a Gaussian prior for g and a log-Gaussian prior for σ,

$$g(x) \sim \mathcal{GP}\big(\mu_g(x), k_\theta^g(x, x')\big), \tag{8}$$

$$\log \sigma(x) \sim \mathcal{GP}\big(\mu_\sigma(x), k_\theta^\sigma(x, x')\big). \tag{9}$$

This model can be compared to the Heteroskedastic GP model introduced by Saul et al. (2016), but with a different likelihood function so that the posterior mode corresponds to a quantile or an expectile.

2.2 INFERENCE PROCEDURE

Although one can obtain a reasonable estimate of a mean value using only a handful of samples, inferring quantiles or expectiles tends to require a much larger number of observations, since they require information associated to the tails of the distribution. The inference procedure for the proposed probabilistic model must thus be able to cope with relatively large datasets, with a number of observations in the order of a few thousands to a tens of thousands data points.

A well established method that supports both large datasets and non-conjugate likelihoods is the Sparse Variational GP framework (Titsias, 2009; Hensman et al., 2013). It consists in approximating the intractable or computationally expensive posterior distribution $p(g, \sigma|\mathcal{Y}_n)$ by a distribution $p(g, \sigma|g(Z)=u_g, \sigma(Z)=u_\sigma)$, where $Z \in \mathcal{X}^N$ and u_g, u_σ are N-dimensional random variables:

$$u_g \sim \mathcal{N}(u_g|\mu_g, S_g) \text{ and } u_\sigma \sim \mathcal{N}(u_\sigma|\mu_\sigma, S_\sigma).$$

The parameters Z, μ_g, S_g, μ_σ, S_σ, are referred to as the variational parameters. The Z's are often called *inducing points*. Intuitively, u_g are random variables that act as pseudo-observations at the inducing point locations.

The variational parameters can be optimised jointly with the model parameters (e.g. mean function coefficients or kernel hyperparameters) such that Kullback-Leibler divergence between the approximate and the true posterior is as small as possible. In practice, this is achieved by maximising the Evidence Lower Bound (ELBO):

$$\sum_{i=1}^{n} \int \log p\big(y_i|g_i, \sigma_i\big)\tilde{p}(g_i)\tilde{p}(\sigma_i)dg_i d\sigma_i \\ - \mathrm{kl}\left(\tilde{p}(u_q)||p(u_q)\right) - \mathrm{kl}\left(\tilde{p}(u_\sigma)||p(u_\sigma)\right),$$

where $\tilde{p}(g_i)$ and $\tilde{p}(\sigma_i)$ are shorthands for the variational posterior distributions at x_i:

$$\tilde{p}(g_i) = \int p(g(x_i)|g(Z) = u_g)p(u_g)du_g \\ = \mathcal{N}(g_i|K_{x_i,u_g}K_{u_g,u_g}^{-1}\mu_g, K_{x_i,x_i} + Q_g),$$

where $Q_g = K_{x_i,u_g} K_{u_g,u_g}^{-1} (S_g - K_{u_g,u_g}) K_{u_g,u_g}^{-1} K_{u_g,x_i}$.

This proposed inference scheme is similar to the one used in Saul et al. (2016), with the notable difference that non differentiability of the pinball loss at the origin implies that we need to resort to using a first order optimizer such as ADAM (Kingma and Ba, 2014) that can handle non-differentiability of the objective function.

3 BAYESIAN OPTIMISATION

Classical BO algorithms work as follow. First, a posterior distribution on g is inferred from an initial set of experiments $\mathcal{D}_n$ (typically obtained using a space-filling design). Then the next input point to evaluate is chosen as the maximiser of an *acquisition function* $\alpha_n : \mathcal{X} \to \mathbb{R}$, computed from the posterior. The objective function is sampled at the chosen input and the posterior on g is updated. These steps are repeated until the budget is exhausted. The efficiency of such strategies depends on how informative the g posterior is but also on the exploration/exploitation trade-off provided by the acquisition function. Many acquisition functions have been designed to control this trade off, among them the *Expected improvement* (EI, Jones et al., 1998), *upper confidence bound* (UCB, Srinivas et al., 2009), *knowledge gradient* (KG, Frazier et al., 2009) and *Entropy search* (PES, Hernández-Lobato et al., 2014).

In the case of quantiles and expectiles, adding points one at a time is impractical since many points are typically necessary to modify significantly the g posterior. Hence, we focus here on *batch-BO* strategies, for which the acquisition recommends a batch of $B > 1$ points instead of a single one. The above-mentioned acquisition functions have been extended to handle batches: see for instance Marmin et al. (2015) for EI, (Wu and Frazier, 2016) for KG or Desautels et al. (2014) for UCB. However, none actually fit our settings for two main reasons. First, most parallel acquisitions make use of explicit update equations for the GP moments and assume access to a Gaussian posterior for observations, neither of which are available for our model. Secondly, most are designed for small batches (say, $B \leq 5$) and become numerically intractable for the larger batches (say, $B > 50$) that are more in line with the data volumes necessary for quantiles and expectiles estimation.

We propose in the following the first acquisition functions that can be applied to our quantile GP surrogate model, one based on Thompson sampling and one on entropy search.

3.1 THOMPSON SAMPLING

Thompson sampling (TS) is becoming increasingly popular in BO, in particular because of its embarrassingly parallel nature allowing full scalability with the batch size (Hernández-Lobato et al., 2017; Kandasamy et al., 2018; Vakili et al., 2021).

Given the posterior on g, an intuitive approach is to sample Ψ according to the probability that x is the location of the maximum of g. Despite this distribution usually being intractable, one may achieve the same result by sampling a trajectory from the posterior of g and then selecting the input that corresponds to its maximiser. Such approach directly extends to batches of inputs, by drawing several trajectories and selecting all the maximisers.

The main drawback of GP-based TS is the cost of sampling a trajectory, which can only be done exactly at a finite number of input locations at a cubic cost in the number of locations. An alternative is to rely on a finite rank approximation of the kernel, but this has been found to have an undesirable effect known as *variance starvation* (Wang et al., 2018).

Wilson et al. (2020) showed that pairing sparse GP models with the so-called *decoupled sampling* formulation avoids the variance starvation issue. Vakili et al. (2021) then demonstrated that such an approach delivered excellent empirical performance on high noise, large budget, large batch scenarios, while enjoying the same theoretical guarantees as the vanilla TS approach. Here, we build upon Vakili et al. (2021), and apply their algorithm to the variational posterior of g to obtain draws directly from the quantile or expectile model. The posterior over σ, which controls the observation noise, is not used during the TS algorithm.

The procedure for generating quantile samples from the variational posterior of g can be summarised as follows: First, a continuous sample from the prior of g is generated using Random Fourier Features (see supplementary material B). Second we sample from the inducing variables u_g. Third, we compute the mean function $m(x)$ of a GPR model that interpolates the dataset $\{Z, u_g - s(Z)\}$. Finally, the posterior sample is obtained by correcting the prior samples with the mean function $v(x) = s(x) + m(x)$.

3.2 INFORMATION-THEORETIC QUANTILE OPTIMISATION WITH GIBBON

Another particularly intuitive search strategy for BO is to choose the evaluations that will maximally reduce the uncertainty in the minimiser of the objective, an approach known as max-value (or min-value) entropy search (MES, Wang and Jegelka, 2017). For quantile optimisation, MES corresponds to reducing uncertainty in the minimal quantile value $g^* = \min_{x \in \mathcal{X}} g(x)$. Following the arguments of Wang and Jegelka (2017), a meaningful measure of uncertainty reduction in this context is taken as the gain in mutual information between a set of candidate evaluations and g^* (see Cover and Thomas, 2012, for an introduction to information theory). Principled information-theoretic optimisation then corresponds to finding batches of B input points $\{x_i\}_{i=1}^B$ that maximise

$$\alpha_n(\{x_i\}_{i=1}^B) = \text{MI}(g^*; \{y_{x_i}\}_{i=1}^B | \mathcal{D}_n), \qquad (10)$$

where y_{x_i} are not-yet-observed evaluations of the batch that are estimated with the GP surrogate model.

Although calculating the acquisition function (10) is challenging, there exist effective approximation strategies for GP models with conjugate likelihoods (Moss et al., 2020b; Takeno et al., 2020). In the remaining of this section we show that the approach used in General-purpose Information Based Bayesian-OptimisatioN (GIBBON Moss et al., 2021) can be adapted to support asymmetric Laplace or Gaussian likelihood so that information-theoretic acquisition functions can be used for our quantile and expectile models.

Following the derivations of Moss et al. (2021), the application of three well-known information-theoretic inequalities provides the following lower-bound for the mutual information (10):

$$\text{MI}(g^*; \{y_{x_i}\}_{i=1}^B | \mathcal{D}_n) \geq \text{H}(\{y_{x_i}\}_{i=1}^B | \mathcal{D}_n) - \frac{1}{2} \sum_{i=1}^B \mathbb{E}_{g^* | \mathcal{D}_n} \left[\log(2\pi e \text{Var}(y_{x_i} | g^*, \mathcal{D}_n)) \right], \qquad (11)$$

where $\text{H}(A) = -\mathbb{E}_A [\log p(A)]$ denotes differential entropy. Although calculating the expectation in the second term of (11) is intractable (i.e. no closed-form expression exists for $p(g^* | \mathcal{D}_n)$), we follow another approximation common among information-theoretic acquisition functions and approximate the integral using Monte-Carlo over a set of M sampled minimum values. In particular, we use the Gumbel sampler proposed by Wang and Jegelka (2017), which provides a cheap set of samples $\mathcal{M}_n = \{g_1^*, .., g_M^*\}$ from $p(g^* | \mathcal{D}_n)$.

When calculating the original GIBBON acquisition function, all the terms in the lower bound (11) are tractable, i.e. the conjugancy of their Gaussian likelihood means that $\text{H}(\{y_{x_i}\}_{i=1}^B | \mathcal{D}_n)$ is just the differential entropy of a multivariate Gaussian which, alongside each $\text{Var}(y_{x_i} | g^*, \mathcal{D}_n)$, has a closed-form expression (See Moss et al. (2021) for details). Consequently, this lower bound itself is used as a closed-form approximation to the mutual information. However, in our quantile setting, we no longer have expressions for the first term of (11) — the joint differential entropy of B-dimensional asymmetric Laplace variables with a complex correlation structure given by our two latent GPs.

To build an information-theoretic acquisition function suitable for our quantile model, we must apply an additional approximation. In particular, by using a moment-matching approximation, we can replace the intractable joint differential entropy with the differential entropy of a multivariate Gaussian of the same covariance, leading to our propose Quantile GIBBON (Q-GIBBON) acquisition function

$$\alpha_n^{\text{Q-GIBBON}} = \frac{1}{2} \log |C| - \frac{1}{2M} \sum_{g^* \in \mathcal{M}_n} \sum_{i=1}^B \log V_i(g^*),$$

where $|C|$ is the determinant of the $B \times B$ predictive covariance matrix with elements $C_{i,j} = \text{Cov}(y_{x_i}, y_{x_j})$ and $V(g^*)$ denotes the conditional variances $V_i(g^*) = \text{Var}(y_{x_i} | g^*, \mathcal{D}_n)$. Crucially, all the terms of Q-GIBBON have closed-form expressions (see appendix A for a derivation of C and V from our quantile GP).

Although applying an additional moment-matching approximation means that Q-GIBBON is no longer a lower bound on the true mutual information, we found that it provides very efficient optimisation (see Section 4). In fact, we tried much more expensive but unbiased Monte-Carlo approximations which did not result in noticeable difference in performance.

In practice, directly searching for the set of B points that maximise $\alpha_n^{\text{Q-GIBBON}}$ is a very challenging task, due to the dimensionality ($B \times D$) and multimodality of the acquisition function. However, the Q-GIBBON formulation makes it particularly well-suited for a *greedy* approach, where we first optimise Q-GIBBON for $B = 1$, then optimise for $B = 2$ while fixing the first point to the previously found value, etc. until B points are found.

4 EXPERIMENTS

We now evaluate our proposed model and acquisition functions on a set of synthetic tasks and two real-world optimisation problems. All that follows could equivalently be applied to expectiles, experiments are focused on quantile optimisation to streamline the exposition. The results presented in this section can be replicated using the code available at `www.github.com/obfuscated-url`.

4.1 ALGORITHM BASELINES

To our knowledge, there is no other existing BO algorithm dedicated to optimising quantiles in our considered setting. The most similar algorithms are those of Cakmak et al. (2020) and Makarova et al. (2021). However, Cakmak et al. (2020) requires precise control over the noise generation process, while Makarova et al. (2021) seek to find solutions with low levels of observation noise but do not provide a method for optimising a specific quantile level.

We can, however, apply standard BO methods to perform quantile optimisation if direct observations of the quantiles are available. This is achievable by using repeated observations, which allows computing a (pointwise) empirical quantile. As direct observations are available, a standard GP

Regression model (GPR) can be used to provide a posterior on g (Plumlee and Tuo, 2014). One can also bootstrap the repeated observations to obtain variance estimates of the empirical quantiles, to improve further the model by accounting for varying observation noise. Next, a BO procedure can be defined based on any classical acquisition function. Here we choose the vanilla EI one. With this strategy, each batch consists of a single point in the input space, repeated a number of times. In the following experiments we use this baseline (denoted GPR-EI) to compare with our two proposed methods using TS and Q-GIBBON over a quantile GP.

4.2 IMPLEMENTATION

All models are built using the `gpflux` library (Dutordoir et al., 2021), and the BO procedure is done using `trieste` (Berkeley et al., 2022). All models use a Matern 5/2 kernel, and all acquisition functions (or GP samples in the case of TS) are optimised using a multi-start BFGS scheme.

Our quantile model requires a design choice for the inducing points placement, these are reinitialised for each model fit. We follow the findings of Vakili et al. (2021) and use the centroids of a k-means procedure on the data points, which tends to concentrate the inducing points near the optimal areas as more data is collected by BO. Our implementation of decoupled Thompson sampling uses 1000 random Fourier features (see supplementary material for detailed expressions). To sample minimum values for Q-GIBBON we use the Gumbel sampler of Wang and Jegelka (2017) with $10,000 \times D$ random initial points.

4.3 SYNTHETIC PROBLEMS

Problem description We generated a set of synthetic problems based on the Generalised Lambda Distribution (GLD, Freimer et al., 1988), a highly flexible four-parameter probability distribution function designed to approximate several well-known parametric distributions. The four parameters define the location, scale, left and right shape of the distribution, respectively. By varying the value of each parameter as a function of x, one can create a black-box with high noise, heteroscedasticity and non-Gaussianity:

$$Y_x \sim GLD(\lambda_1(x), \ldots, \lambda_4(x))). \tag{12}$$

To generate a large set of problems with varying dimensionality while controlling the multimodality of the problem at hand, we used GP random draws for the λ_i's. See appendix for a full description. Figure 2 shows examples of marginal distributions (for different x values) for one such problem.

We consider two input space dimensions: $D = 3$ and 6 and two quantile levels, $\tau = 0.75$ and 0.95. We use as an initial budget $50D$ observations, uniformly distributed across the input space and a total budget of $250D$ observations, acquired in batches of either $B = 10$ or 50 points. Each strategy is run on 50 different problems. We report here the simple regret in Figure 3, averaged over the 50 problems, with confidence intervals.

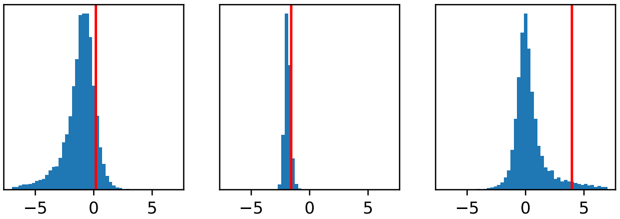

Figure 2: Examples of marginal distributions for one GLD-based problem at three different locations of the input space.

Results In almost all cases, our approaches largely outperform the GPR baseline, the exception being on the simpler problem (small dimension and batch size) for which the GPR baseline is comparable to TS (GIBBON being substantially better for the 0.75 quantile). Comparing acquisition strategies, GIBBON clearly outperforms TS for $D = 3$. In dimension 6, both approaches are roughly comparable.

4.4 LUNAR LANDER

Problem description The Lunar Lander problem is a popular benchmark for noisy BO (Moss et al., 2020a; Eriksson et al., 2019). In this well-known reinforcement learning task, we must control three engines (left, main and right) to successfully land a rocket. The learning environment and a hard-coded PID controller is provided in the OpenAI gym.[1] We seek to optimise 6 thresholds present in the description of the controller to provide the largest expected reward: finding those thresholds defines the BO task. Our RL environment is exactly as provided by OpenAI. We lose 0.3 points per second of fuel use and 100 if we crash. We gain 10 points each time a leg makes contact with the ground, 100 points for any successful landing, and 200 points for a successful landing in the specified landing zone. Each individual run of the environment allows the testing of a controller on a specific random seed.

This problem is particularly well-suited for a quantile approach, since reward is stochastic, highly non-Gaussian, and the landing problem is a clear case for which one would want guarantees against risk.

Results For this problem, we ran each algorithm 10 times (starting from different initial conditions), with batches of $B = 25$ points 300 initial observations and 1500 in total. We aim to maximise the 10% quantile of the reward. Due to the high cost of calculating the true quantiles of the lunar

[1] *https://gym.openai.com/*

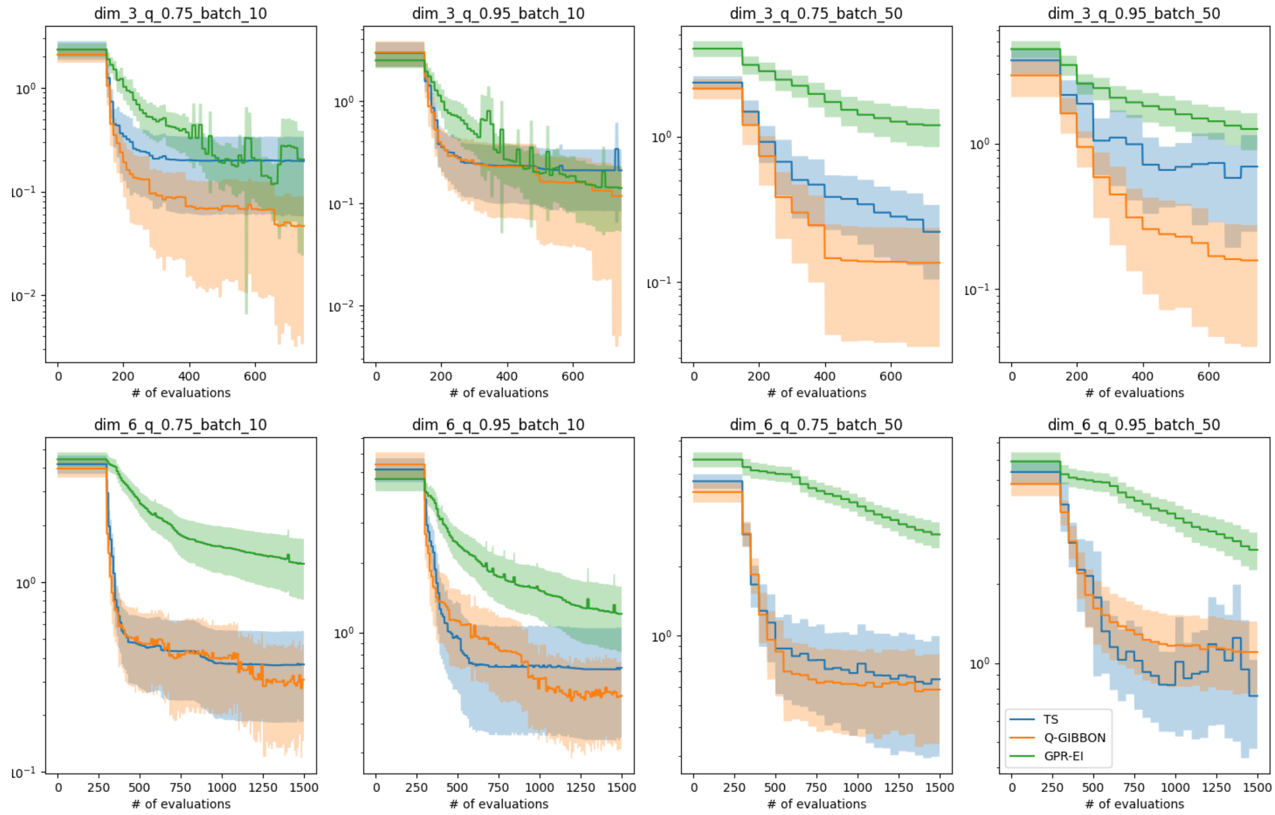

Figure 3: The mean and 95% confidence intervals of regret on synthetic problems in dimension 3 (top) and 6 (bottom), for two quantile levels ($\tau = 0.75, 0.95$) and medium ($B = 10$, left) and large ($B = 50$, right) batch sizes.

lander experiment (i.e. they must be calculated empirically across a large collection of runs), we only report the reward quantile obtained after half and all the iterations (see Table 1) and only run one of our two proposed acquisition functions. We choose TS over GIBBON as our synthetic GLD experiments suggest that TS outperforms Q-GIBBON on problems with larger (i.e. 6) dimensions. We can see that TS largely outperforms the baseline, as it seems to robustly identify a much better solution.

	750 obs	1500 obs
GPR-EI	94.6 (106.1)	159.5 (110.9)
TS	204.3 (53.8)	255.2 (8.0)

Table 1: Mean and standard deviation over 10 runs for the 10% quantile of the reward on the lunar lander problem.

4.5 LASER TUNING

Problem Description For our final experiment, we test our quantile optimisation in a real-world setting inspired by the Free-Electron Laser (FEL) tuning example of McIntire et al. (2016). This is a challenging 16-dimensional optimisation task where we must configure the strengths of magnets manipulating the shape of the FEL's electron beam, seeking to build a powerful and stable beam suitable for use in scientific experiments. Due to the high levels of observation noise in this problem and as stability of the resulting beam is of critical importance for conducting reliable experiments, it is clearly beneficial to encode a level of risk-adversity into the optimisation. Therefore, there are clear advantages for using quantile optimisation for FEL calibration.

As we do not have access to the FEL directly, we follow McIntire et al. (2016) and use their $4,074$ observed X-ray pulse energy measurements to build Gaussian process surrogate model from which we can simulate pulse energy at any new magnet configuration. To simulate the effect of observation noise, McIntire et al. (2016) add additional Gaussian perturbations to the simulated values. However, we found that the noise in this system was actually skew Gaussian and varied in scale and skew across the search space. Consequently, we simulate observation noise from a skew Gaussian distribution with location, scale and shape parameters also modelled with additional GPs (i.e. a setup similar to our GLD examples). As many of the $4,074$ energy measurements are evaluated at very similar input locations, rounding these inputs to four decimal places provides us with many repeated evaluations, allowing the empirical estimation of each parameter of the skew Gaussian distribution at each of these inputs. The location, scale and shape GPs

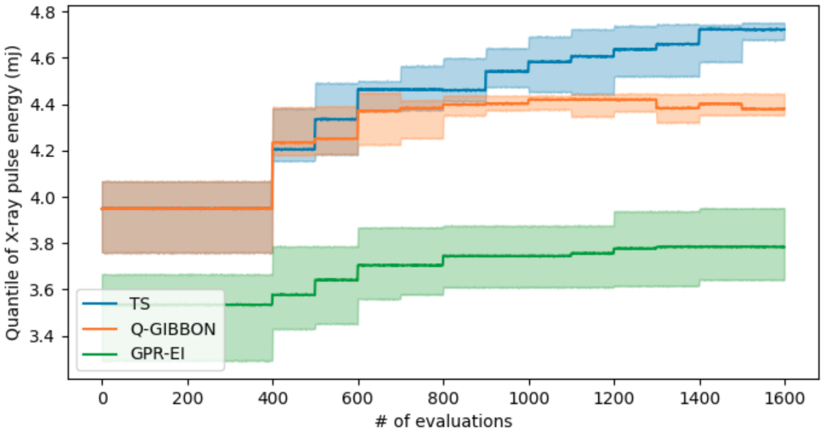

Figure 4: The mean and 95% confidence intervals of best 0.3 quantile found across 10 repetitions of the FEL tuning task.

are then determined to predict the parameters of the skew Gaussian noise distributions for any candidate magnet configuration.

Results Figure 4 shows the performance of each algorithm over 10 repetitions, seeking to maximise the 30% quantile of pulse energy. The models are initialised with 400 data points randomly chosen from the full dataset, and a further 1,200 points are collected with BO in batches of 100 points. Our algorithms based on quantile GP models substantially outperform the replicate-based GPR baseline. In fact, by using TS with a quantile GP, we are able to find solutions very close to the optimal value (4.8). We hypothesise that the relatively poor performance of our Q-GIBBON acquisition function is due to the high dimension of this problem. The Gumbel sampler used by Q-GIBBON for sampling minimal-values is based on random sampling and so its performance likely degrades as the input dimension increases. Since the performance of information-theoretic BO is sensitive to the quality of these samples (Moss et al., 2021), extending information-theoretic BO to high dimensional problems like FEL tuning remains an open question.

4.6 CONCLUDING COMMENTS

We have presented a new setting to estimate quantiles and expectiles of stochastic black box functions that is well suited to heteroscedastic cases. We then used the proposed model to create two BO algorithms designed for the optimisation of conditional quantiles and expectiles without repetitions in the experimental design. These algorithms outperform the state of the art on several test problems with different dimensions, quantile orders, budgets and batch sizes.

Overall, our experiments clearly show that the performance gap between our approaches and the GPR-EI baseline increases with the batch size and problem dimension. Since GPR-EI relies on repetitions, it is much more limited in terms of exploration, while our approaches can evaluate B unique points at each BO iteration. Hence, our approach is much less sensitive to the curse of dimensionality.

Experiments also show that for low-dimensional, smaller batches, Q-GIBBON is the best alternative, while with increasing dimension and batch size, the simpler Thompson sampling seems to perform best. Depending on the available hardware, the parallel nature of TS might also provide substantial advantages in terms of wall-clock time.

Bibliography

Abeywardana, S. and Ramos, F. (2015). Variational inference for nonparametric bayesian quantile regression. In AAAI, pages 1686–1692.

Bellini, F. and Di Bernardino, E. (2017). Risk management with expectiles. The European Journal of Finance, 23(6):487–506.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Advances in neural information processing systems, pages 2546–2554.

Berkeley, J., Moss, H. B., Artemev, A., Pascual-Diaz, S., Granta, U., Stojic, H., Couckuyt, I., Qing, J., Loka, N., Paleyes, A., Ober, S. W., and Picheny, V. (2022). Trieste v.0.10.0. https://github.com/secondmind-labs/trieste.

Boukouvalas, A., Barillec, R., and Cornford, D. (2012). Gaussian process quantile regression using expectation propagation. arXiv preprint arXiv:1206.6391.

Browne, T., Iooss, B., Gratiet, L. L., Lonchampt, J., and Remy, E. (2016). Stochastic simulators based optimization by gaussian process metamodels–application to maintenance investments planning issues. Quality and Reliability Engineering International, 32(6):2067–2080.

Cakmak, S., Astudillo Marban, R., Frazier, P., and Zhou, E. (2020). Bayesian optimization of risk measures. Advances in Neural Information Processing Systems, 33:20130–20141.

Cannon, A. J. (2011). Quantile regression neural networks: Implementation in r and application to precipitation downscaling. Computers & geosciences, 37(9):1277–1284.

Cover, T. M. and Thomas, J. A. (2012). Elements of information theory. John Wiley & Sons.

Desautels, T., Krause, A., and Burdick, J. W. (2014). Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. The Journal of Machine Learning Research, 15(1):3873–3923.

Dutordoir, V., Salimbeni, H., Hambro, E., McLeod, J., Leibfried, F., Artemev, A., van der Wilk, M., Hensman, J., Deisenroth, M. P., and John, S. (2021). Gpflux: A library for deep gaussian processes. arXiv preprint arXiv:2104.05674.

Eriksson, D., Pearce, M., Gardner, J., Turner, R. D., and Poloczek, M. (2019). Scalable global optimization via local bayesian optimization. Advances in Neural Information Processing Systems, 32.

Farooq, M. and Steinwart, I. (2017). An svm-like approach for expectile regression. Computational Statistics & Data Analysis, 109:159–181.

Frazier, P., Powell, W., and Dayanik, S. (2009). The knowledge-gradient policy for correlated normal beliefs. INFORMS journal on Computing, 21(4):599–613.

Freimer, M., Kollia, G., Mudholkar, G. S., and Lin, C. T. (1988). A study of the generalized Tukey lambda family. Communications in Statistics-Theory and Methods, 17(10):3547–3567.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. arXiv preprint arXiv:1309.6835.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. In Advances in neural information processing systems, pages 918–926.

Hernández-Lobato, J. M., Requeima, J., Pyzer-Knapp, E. O., and Aspuru-Guzik, A. (2017). Parallel and distributed thompson sampling for large-scale accelerated exploration of chemical space. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1470–1479. JMLR. org.

Jiang, C., Jiang, M., Xu, Q., and Huang, X. (2017). Expectile regression neural network model with applications. Neurocomputing, 247:73–86.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. Journal of Global optimization, 13(4):455–492.

Kandasamy, K., Krishnamurthy, A., Schneider, J., and Póczos, B. (2018). Parallelised bayesian optimisation via thompson sampling. In International Conference on Artificial Intelligence and Statistics, pages 133–142. PMLR.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. Econometrica: journal of the Econometric Society, pages 33–50.

Makarova, A., Usmanova, I., Bogunovic, I., and Krause, A. (2021). Risk-averse heteroscedastic bayesian optimization. Advances in Neural Information Processing Systems, 34.

Marmin, S., Chevalier, C., and Ginsbourger, D. (2015). Differentiating the multipoint expected improvement for optimal batch design. In International Workshop on Machine Learning, Optimization and Big Data, pages 37–48. Springer.

McIntire, M., Ratner, D., and Ermon, S. (2016). Sparse gaussian processes for bayesian optimization. In UAI.

Meinshausen, N. (2006). Quantile regression forests. Journal of Machine Learning Research, 7(Jun):983–999.

Mockus, J., Tiesis, V., and Zilinskas, A. (1978). The application of bayesian methods for seeking the extremum. Towards global optimization, 2(117-129):2.

Moss, H. B., Leslie, D. S., Gonzalez, J., and Rayson, P. (2021). Gibbon: General-purpose information-based bayesian optimisation. Journal of Machine Learning Research, 22(235):1–49.

Moss, H. B., Leslie, D. S., and Rayson, P. (2020a). Bosh: Bayesian optimization by sampling hierarchically. arXiv preprint arXiv:2007.00939.

Moss, H. B., Leslie, D. S., and Rayson, P. (2020b). Mumbo: Multi-task max-value bayesian optimization. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 447–462. Springer.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. Econometrica: Journal of the Econometric Society, pages 819–847.

Picheny, V., Wagner, T., and Ginsbourger, D. (2013). A benchmark of kriging-based infill criteria for noisy optimization. Structural and Multidisciplinary Optimization, 48(3):607–626.

Plumlee, M. and Tuo, R. (2014). Building accurate emulators for stochastic simulations via quantile kriging. Technometrics, 56(4):466–473.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In Summer School on Machine Learning, pages 63–71. Springer.

Rockafellar, R. T., Uryasev, S., et al. (2000). Optimization of conditional value-at-risk. Journal of risk, 2:21–42.

Rostek, M. (2010). Quantile maximization in decision theory. The Review of Economic Studies, 77(1):339–371.

Saul, A. D., Hensman, J., Vehtari, A., and Lawrence, N. D. (2016). Chained gaussian processes. In Artificial Intelligence and Statistics, pages 1431–1440.

Skullerud, H. (1968). The stochastic computer simulation of ion motion in a gas subjected to a constant electric field. Journal of Physics D: Applied Physics, 1(11):1567.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv preprint arXiv:0912.3995.

Székely Jr, T. and Burrage, K. (2014). Stochastic simulation in systems biology. Computational and structural biotechnology journal, 12(20-21):14–25.

Takeno, S., Fukuoka, H., Tsukada, Y., Koyama, T., Shiga, M., Takeuchi, I., and Karasuyama, M. (2020). Multi-fidelity bayesian optimization with max-value entropy search and its parallelization. In International Conference on Machine Learning, pages 9334–9345. PMLR.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric quantile estimation. Journal of machine learning research, 7(Jul):1231–1264.

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In Artificial Intelligence and Statistics, pages 567–574.

Torossian, L., Picheny, V., Faivre, R., and Garivier, A. (2019). A review on quantile regression for stochastic computer experiments. arXiv preprint arXiv:1901.07874.

Vakili, S., Moss, H., Artemev, A., and Picheny, V. (2021). Scalable thompson sampling using sparse gaussian process models. In Advances in neural information processing systems. PMLR.

Wang, Z., Gehring, C., Kohli, P., and Jegelka, S. (2018). Batched large-scale bayesian optimization in high-dimensional spaces. In International Conference on Artificial Intelligence and Statistics, pages 745–754. PMLR.

Wang, Z. and Jegelka, S. (2017). Max-value entropy search for efficient bayesian optimization. In International Conference on Machine Learning, pages 3627–3635. PMLR.

Wilson, J., Borovitskiy, V., Terenin, A., Mostowsky, P., and Deisenroth, M. (2020). Efficiently sampling functions from gaussian process posteriors. In International Conference on Machine Learning, pages 10292–10302. PMLR.

Wu, J. and Frazier, P. (2016). The parallel knowledge gradient method for batch bayesian optimization. In Advances in Neural Information Processing Systems, pages 3126–3134.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. Statistics & Probability Letters, 54(4):437–447.

A SUPPLEMENTARY MATERIAL: CALCULATION OF Q-GIBBON

We derive here the analytical form of our proposed Q-GIBBON acquisition function. For simplicity, we focus on the quantile setting, but the expectile case only requires a straightforward modification of the following derivation.

Recall that Q-GIBBON is defined as

$$\alpha_n^{\text{Q-GIBBON}} = \frac{1}{2}\log|C| - \frac{1}{2M}\sum_{g^*\in\mathcal{M}_n}\sum_{i=1}^{B}\log V_i(g^*),$$

where $|C|$ is the determinant of the $B \times B$ predictive covariance matrix with elements $C_{i,j} = \text{Cov}(y_{x_i}, y_{x_j}|\mathcal{D}_n)$ and $V(g^*)$ denotes the conditional variances $V_i(g^*) = \text{Var}(y_{x_i}|g^*, \mathcal{D}_n)$. Therefore, calculating Q-GIBBON boils down to being able to calculate $V_i(g^*)$ and $C_{i,j}$ across any candidate batch of points (i.e. for all $i, j \in \{1, .., B\}$). We now derive closed-form expressions for $V_i(g^*)$ and $C_{i,j}$.

A.1 REQUIRED PREDICTIVE QUANTITIES

For ease of notation, we will consider just a single pair of input values of x_1 and x_2 and show how to calculate $V_1(g^*)$ and $C_{1,2}$. Denote the quantiles, scales and (noisy) observations at these two location as $g_1 = g(x_1)|\mathcal{D}_n$, $g_2 = g(x_2)|\mathcal{D}_n$, $\sigma_1 = \sigma(x_1)|\mathcal{D}_n$, $\sigma_2 = \sigma(x_2)|\mathcal{D}_n$, $y_1 = y(x_1)|\mathcal{D}_n$ and $y_2 = y(x_2)|\mathcal{D}_n$, respectively. Then, from our underlying GP models we can extract our current beliefs about these random variables:

$$\begin{aligned}\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} &\sim N\left[\begin{pmatrix} \mu_1^g \\ \mu_2^g \end{pmatrix}, \begin{pmatrix} (\sigma_1^g)^2 & \Sigma_{1,2}^g \\ \Sigma_{1,2}^g & (\sigma_2^g)^2 \end{pmatrix}\right], \\ \begin{pmatrix} \log(\sigma_1) \\ \log(\sigma_2) \end{pmatrix} &\sim N\left[\begin{pmatrix} \mu_1^\sigma \\ \mu_2^\sigma \end{pmatrix}, \begin{pmatrix} (\sigma_1^\sigma)^2 & \Sigma_{1,2}^\sigma \\ \Sigma_{1,2}^\sigma & (\sigma_2^\sigma)^2 \end{pmatrix}\right].\end{aligned}$$

For closed form expressions of μ_1^g, σ_1^g, ... see any GP textbook, e.g. Rasmussen (2003).

Before deriving expressions for $V_1(g^*)$ and $C_{1,2}$, it is convenient to write the conditional mean and variance of our noisy observations y_1 and y_2. Following Yu and Moyeed (2001), we have

$$\mathbb{E}[y_1|g_1,\sigma_1] = g_1 + \frac{1-2\tau}{\tau(1-\tau)}\sigma_1, \tag{13}$$

$$\text{Var}(y_1|g_1,\sigma_1) = \frac{1-2\tau+2\tau^2}{\tau^2(1-\tau)^2}\sigma_1^2, \tag{14}$$

with similar expressions for the moments of $y_2|g_2,\sigma_2$

A.2 CALCULATING THE CONDITIONAL VARIANCE V

We now have all the quantities required to calculate $V_1(g^*) = \text{Var}(y|g^*)$. Recall that g^* denotes the maximal value obtained by the quantile (i.e. $g(x)$). First, we use the law of total variance to decompose V_1 into two terms:

$$\begin{aligned}V_1 =& \text{Var}_{g_1,\sigma|g^*}\left(\mathbb{E}[y_1|g_1,\sigma_1,g^*]\right) \\ &+ \mathbb{E}_{g_1,\sigma|g^*}\left[\text{Var}\left(y_1|g_1,\sigma_1,g^*\right)\right].\end{aligned} \tag{15}$$

Note that conditioning on g_1, σ, g^* is equivalent to conditioning on g_1, σ only, as knowing that $g^* = \min g(x)$ does not provide additional information over knowing g_1 itself. Therefore, we can insert our expressions for the moments of the asymmetric Laplace (13) and (14) into (15) which, after simple manipulation provides:

$$\begin{aligned}V_1(g^*) = \text{Var}_{g_1|g^*}(g_1) &+ \frac{3(1-2\tau)^2+1}{2\tau^2(1-\tau)^2}\mathrm{e}^{2(\mu_1^\sigma+(\sigma_1^\sigma)^2)} \\ &+ \frac{(1-2\tau)^2}{2\tau^2(1-\tau)^2}\mathrm{e}^{2\mu_1^\sigma+(\sigma_1^\sigma)^2}.\end{aligned} \tag{16}$$

All that remains for the calculation of $V(g^*)_1$ is an expression for $\text{Var}_{g_1|g^*}(g_1)$. Fortunately, as shown by Wang and Jegelka (2017), $g|g^*$ is simply an upper truncated Gaussian variable. Therefore, using the well-known expression for the variance of a truncated Gaussian, we have

$$\text{Var}_{g_1|g^*}(g_1) = (\sigma_1^g)^2\left(1 + \frac{\phi(\gamma_{g^*})}{\Psi(\gamma_{g^*})}\left(\gamma_{g^*} - \frac{\phi(\gamma_{g^*})}{\Psi(\gamma_{g^*})}\right)\right), \tag{17}$$

where $\gamma_{g^*} = \frac{g^*-\mu_1^g}{\sigma_1^g}$, and ϕ and Ψ are the probability density functions and cumulative density functions of a standard Gaussian variable, respectively.

Finally, inserting (17) into (16) yields a closed form expression for $V_1(g^*)$.

A.3 CALCULATING THE PREDICTIVE COVARIANCE C

Just like when calculating the conditional variance V_1, we begin our decomposition of $C_{1,2} = Cov(y_1, y_2)$ by applying the law of total variance to get the following two term expansion:

$$\begin{aligned}C_{1,2} =& \text{Cov}_{g_1,g_2,\sigma_1,\sigma_2}\left(\mathbb{E}\left[y_1|g_1,\sigma_1\right], \mathbb{E}\left[y_2,g_2,\sigma_2\right]\right) \\ &+ \mathbb{E}_{g_1,g_2,\sigma_1,\sigma_2}\left[\text{Cov}(y_1,y_2|g_1,g_2,\sigma_1,\sigma_2)\right].\end{aligned} \tag{18}$$

Now, as $y_1|g_1,\sigma_1$ and $y_2|g_2,\sigma_2$ are independent (all that remains after this conditioning is observation noise), the second term of (18) is in fact zero (at least for unique x_1 and x_2).

To calculate the first term of (18), we insert the expression for the first moment of $y|g, \sigma$ (i.e. Equation (13)) which, after recalling the independence of g and σ, yields

$$\begin{aligned} C_{1,2} = &\, \mathrm{Cov}_{g_1,g_2}(g_1, g_2) \\ &+ \frac{(1-2\tau)^2}{\tau^2(1-\tau)^2}\mathrm{Cov}_{\sigma_1,\sigma_2}(\sigma_1, \sigma_2). \end{aligned} \tag{19}$$

Finally, we can extract $\mathrm{Cov}(g_1, g_2)$ and $\mathrm{Cov}(\sigma_1, \sigma_2)$ from our underlying GP models as $\Sigma^g_{1,2}$ and $\mathrm{e}^{\mu^\sigma_1+\mu^\sigma_2+0.5(\sigma^\sigma_1+\sigma^\sigma_2)}(\mathrm{e}^{\Sigma^\sigma_{1,2}} - 1)$ (using the formulae for the covariance of joint log Gaussian variables). Inserting these two covariances into (19) provides a closed-from expression for $C_{1,2}$.

B SUPPLEMENTARY MATERIAL: RFF FOR MATERN KERNELS

We present in this section how to use RFFs to generate samples from d-dimensional Matern kernels with regularity ν, variance σ^2 and lengthscales $\theta \in \mathbb{R}^d$. First of all, we start from the spectral density of a Matérn kernel:

$$s(w) = \sigma^2|\Lambda|^{1/2}\frac{\Gamma(\frac{d}{2}+\nu)}{\Gamma(\nu)}\frac{(2\sqrt{\pi})^d}{(1+w^T\Lambda w)^{\frac{d}{2}+\nu}},$$

where $\Lambda = \mathrm{diag}(\theta_1, \cdots, \theta_d)$ is the diagonal matrix containing the length scale hyperparameters. Using the change of variable $\Lambda' = 2\nu \times \Lambda$ and introducing rescaling factor $\sigma^2(\sqrt{2}\pi)^d$, one can recognise here the probability density function of the *multivariate t-distribution*:

$$p(w) = |\Lambda|^{1/2}\frac{\Gamma(\frac{d}{2}+\nu)}{\Gamma(\nu)\pi^{d/2}\nu^{d/2}}\frac{1}{(1+\frac{1}{2\nu}w^T\Lambda w)^{\frac{d}{2}+\nu}}.$$

As a consequence, prior samples can be generated by computing

$$g(x) = \sigma\sqrt{2(\sqrt{2}\pi)^d/m}\sum_{i=1}^{m}\omega_i\cos(w_i^Tx+b_i)$$

where $\omega_i \sim \mathcal{N}(0, 1)$, $w_i \sim p$, $b_i \sim \mathcal{U}(0, 2\pi)$, and m is the number of features.

C SUPPLEMENTARY MATERIAL: DESCRIPTION OF THE GLD SYNTHETIC CASE

Several formulations of the GLD exist, we use here the parameterisation of Freimer et al. (1988). The GLD is defined by its quantile function:

$$Q(u) = \lambda_0 + \lambda_1\left(T_1 - T_2\right), \tag{20}$$

with:

$$\begin{aligned} T_1 &= \begin{cases} \frac{u^{\lambda_2}-1}{\lambda_2} & \textit{if } \lambda_2 \neq 0 \\ \log(u) & \textit{if } \lambda_2 = 0 \end{cases} \\ T_2 &= \begin{cases} \frac{(1-u)^{\lambda_3}-1}{\lambda_3} & \textit{if } \lambda_3 \neq 0 \\ \log(1-u) & \textit{if } \lambda_3 = 0 \end{cases} . \end{aligned}$$

Here, the only constraint for the parameter values is $\lambda_1 > 0$.

To define an experiment, each λ_j is a realisation of a GP, except for λ_1 for which we use a softplus transform to ensure positivity:

$$\begin{aligned} \lambda_j(x) &\sim \mathcal{GP}\big(0, k(\cdot,\cdot)\big), \quad j \in \{0, 2, 3\}, \\ \phi(\lambda_1(x)) &\sim \mathcal{GP}\big(0, k(\cdot,\cdot)\big), \end{aligned}$$

with $\phi^{-1}(w) = \log(1 + e^w)$. All GPs have a Matern 5/2 kernel k with unit variance. We add to $\lambda_0(x)$ a small quadratic mean function to avoid having the optimum located on the edges of the domain. We use a lengthscale of 0.5 in dimension 3 and 1.0 in dimension 6. These settings ensure that the 6-dimensional test cases do not have too many local optima.