Sharp Calibrated Gaussian Processes - Supplementary Material

Proof of Theorem 5.4

For completeness, we state Theorem 1 from Marx et al. (2022) here in adapted form, which we then use to prove Theorem 5.4.

Lemma 8.1 (Marx et al. (2022), Theorem 1)**.** *Let $\varphi : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ be a function such that that $\varphi(\boldsymbol{x}, y)$, $\boldsymbol{x}, y \sim \Pi$ is an absolutely continuous random variable and, for any fixed $\boldsymbol{x}^* \in \mathcal{X}$, $\varphi(\boldsymbol{x}^*, \cdot)$ is strictly monotonically increasing. Furthermore, for a set of calibration data $\mathcal{D}_{cal} = \{\boldsymbol{x}_{cal}^i, y_{cal}^i\}$ with $N_{cal} = |\mathcal{D}_{cal}|$ and a permutation $i_1, \ldots, i_{N_{cal}} \in [1, 2, \ldots, N_{cal}]$ such that*

$$\varphi(\boldsymbol{x}_{cal}^{i_j}, y_{cal}^{i_j}) < \varphi(\boldsymbol{x}_{cal}^{i_{j+1}}, y_{cal}^{i_{j+1}}),$$

let $H : \mathbb{R} \to [0, 1]$ be a monotonically non-decreasing function, such that $H(\varphi(\boldsymbol{x}_{cal}^{i_j}, y_{cal}^{i_j})) = \frac{j}{N_{cal}+1}$ holds for all $j = 1, \ldots, N_{cal}$. Then

$$\mathbb{P}_{\boldsymbol{x},y\sim\Pi}\Big(H(\varphi(\boldsymbol{x}, y)) \leq \delta\Big) \in \left[\delta - \frac{1}{N_{cal}+1}, \delta + \frac{1}{N_{cal}+1}\right] \quad \forall\, \delta \in [0, 1].$$

The idea behind the proof of Theorem 5.4 is to show that the solution $\varphi(\boldsymbol{x}, y)$ of the implicit equation

$$y - \mu_{\mathcal{D}_{\text{tr}}}(\boldsymbol{\theta}^R, \boldsymbol{x}) - \hat{\beta}(\varphi(\boldsymbol{x}, y))\sigma\left(\hat{\boldsymbol{\theta}}\left(\varphi(\boldsymbol{x}, y)\right), \boldsymbol{x}\right) = 0 \tag{10}$$

satisfies the requirements stipulated by Lemma 8.1, where $\tilde{\beta}(\delta)$ and $\tilde{\boldsymbol{\theta}}(\delta)$ are *arbitrary* continuous functions such that

$$\begin{aligned} \lim_{\delta\to\infty} \tilde{\beta}(\delta) &= \infty, \qquad \lim_{\delta\to-\infty} \tilde{\beta}(\delta) = -\infty, \\ \lim_{\delta\to\infty} \tilde{\boldsymbol{\theta}}(\delta) &= \infty, \qquad \lim_{\delta\to-\infty} \tilde{\boldsymbol{\theta}}(\delta) = \infty, \end{aligned} \tag{11}$$

$$\begin{aligned} &\tilde{\beta}(\delta) \text{ is strictly monotonically increasing for all } \delta \in \mathbb{R} \\ &\tilde{\boldsymbol{\theta}}(\delta) \text{ is monotonically increasing for all } \delta \in \{\delta \in \mathbb{R} \mid \tilde{\beta}(\delta) > 0\} \\ &\tilde{\boldsymbol{\theta}}(\delta) \text{ is monotonically decreasing for all } \delta \in \{\delta \in \mathbb{R} \mid \tilde{\beta}(\delta) < 0\}. \end{aligned} \tag{12}$$

$\tilde{\boldsymbol{\theta}}(\delta)$ is monotonically increasing with respect to δ for all $\delta \in \{\delta \in \mathbb{R} \mid \tilde{\beta}(\delta) > 0\}$, and monotonically decreasing with respect to δ for all $\delta \in \{\delta \in \mathbb{R} \mid \tilde{\beta}(\delta) < 0\}$. Note that the functions $\hat{\beta}(\delta)$ and $\hat{\boldsymbol{\theta}}(\delta)$ can be easily extended within the real axis to satisfy the requirements mentioned above, which means that they are contained within the set from which $\tilde{\beta}(\delta)$ and $\tilde{\boldsymbol{\theta}}(\delta)$. The reason why we choose arbitrary $\tilde{\beta}(\delta)$ and $\tilde{\boldsymbol{\theta}}(\delta)$, as opposed to the functions $\hat{\beta}(\delta)$ and $\hat{\boldsymbol{\theta}}(\delta)$, is because we need $\varphi(\boldsymbol{x}, y)$ to be independent of the calibration data $\mathcal{D}_{\text{cal}}$ in order to be able to employ Lemma 8.1. Showing that $\varphi(\boldsymbol{x}, y)$ satisfies the requirements of Lemma 8.1 for any $\tilde{\beta}(\delta)$ and $\tilde{\boldsymbol{\theta}}(\delta)$ then implies that we can also choose any function within this class that minimizes sharpness, meaning that these properties also extend to $\hat{\beta}(\delta)$ and $\hat{\boldsymbol{\theta}}(\delta)$.

To prove Theorem 5.4, we will require the following result.

Lemma 8.2. *Consider the regressor $\mu_{\mathcal{D}_{tr}}(\boldsymbol{\theta}^R, \cdot)$, and let $\tilde{\beta}(\delta)$ and $\tilde{\boldsymbol{\theta}}(\delta)$ be functions that satisfy* (11) *and* (12)*. Then, for arbitrary fixed y and $\boldsymbol{x}$,*

$$y - \mu_{\mathcal{D}_{tr}}(\boldsymbol{\theta}^R, \boldsymbol{x}) - \tilde{\beta}(\delta)\sigma\left(\tilde{\boldsymbol{\theta}}\left(\delta\right), \boldsymbol{x}\right) \tag{13}$$

is strictly monotonically decreasing with δ.

Proof. The proof follows directly from Assumption 4.1 and the properties (11) and (12). □

Proof of Theorem 5.4. Let $\tilde{\beta}(\delta)$ and $\tilde{\boldsymbol{\theta}}(\delta)$ be functions that satisfy (11) and (12). Due to Lemma 8.2, we can define the function $\varphi : \mathcal{X} \times \mathbb{R} \to [0, 1]$ as the unique solution to the implicit equation

$$y - \mu_{\mathcal{D}_{\text{tr}}}(\boldsymbol{\theta}^R, \boldsymbol{x}) - \tilde{\beta}(\varphi(\boldsymbol{x}, y))\sigma\left(\tilde{\boldsymbol{\theta}}\left(\varphi(\boldsymbol{x}, y)\right), \boldsymbol{x}\right) = 0. \tag{14}$$

Note that, since $y - \mu_{\mathcal{D}_{\text{tr}}}(\boldsymbol{\theta}^R, \boldsymbol{x})$ is strictly monotonically increasing with y, $\varphi(\boldsymbol{x}, y)$ is a strictly monotonically increasing function of y for any fixed $\boldsymbol{x}$. Furthermore, since y is absolutely continuous,

$$y^i_{\text{cal}} - \mu_{\mathcal{D}_{\text{tr}}}(\boldsymbol{\theta}^R, \boldsymbol{x}^i_{\text{cal}}) \neq y^j_{\text{cal}} - \mu_{\mathcal{D}_{\text{tr}}}(\boldsymbol{\theta}^R, \boldsymbol{x}^j_{\text{cal}})$$

holds for all $i \neq j$ almost surely, which implies $\varphi(\boldsymbol{x}^i_{\text{cal}}, y^i_{\text{cal}}) \neq \varphi(\boldsymbol{x}^j_{\text{cal}}, y^j_{\text{cal}})$ for all $i \neq j$ almost surely. Hence, $\varphi(\boldsymbol{x}, y)$, $\boldsymbol{x}, y \sim \Pi$, corresponds to an absolutely continuous random variable. Hence, given any monotonically non-decreasing function $H(\cdot)$ that satisfies the requirement

$$H\left(\varphi(\boldsymbol{x}^{i_j}_{\text{cal}}, y^{i_j}_{\text{cal}})\right) = \frac{j}{N_{\text{cal}} + 1},$$

Lemma 8.1 implies that

$$\mathbb{P}_{\boldsymbol{x}, y \sim \Pi}\Big(H\left(\varphi(\boldsymbol{x}, y)\right) \leq \delta\Big) \in \left[\delta - \frac{1}{N_{\text{cal}} + 1}, \delta + \frac{1}{N_{\text{cal}} + 1}\right] \quad \forall\, \delta \in [0, 1]. \tag{15}$$

Since $\tilde{\beta}(\delta)$ and $\hat{\boldsymbol{\theta}}\left(\delta\right)$ are arbitrary, and $\hat{\beta}(\delta)$ and $\tilde{\boldsymbol{\theta}}\left(\delta\right)$ are continuous and also satisfy (11) and (12) within $\delta \in [0, 1]$, we can substitute $\varphi(\cdot, \cdot)$ in (15) with $\hat{\varphi}(\cdot, \cdot)$, which is the unique solution of the implicit equation

$$y - \mu_{\mathcal{D}_{\text{tr}}}(\boldsymbol{\theta}^R, \boldsymbol{x}) - \hat{\beta}(\hat{\varphi}(\boldsymbol{x}, y))\sigma\left(\hat{\boldsymbol{\theta}}\left(\hat{\varphi}(\boldsymbol{x}, y)\right), \boldsymbol{x}\right) = 0. \tag{16}$$

Now, in the particular case of $\hat{\varphi}(\cdot, \cdot)$, due to (7), we have that

$$\hat{\varphi}(\boldsymbol{x}^{i_j}_{\text{cal}}, y^{i_j}_{\text{cal}}) = \frac{j}{N_{\text{cal}} + 1},$$

meaning that $H(\hat{\varphi}(\boldsymbol{x}^{i_j}_{\text{cal}}, y^{i_j}_{\text{cal}})) = \hat{\varphi}(\boldsymbol{x}^{i_j}_{\text{cal}}, y^{i_j}_{\text{cal}})$, i.e., (15) holds for $\varphi(\cdot, \cdot) = \hat{\varphi}(\cdot, \cdot)$ and the identity function $H(\delta) = \delta$. Furthermore, since $\hat{\varphi}(\cdot, \cdot)$ is uniquely defined by the implicit equation (16) and $\hat{\beta}(\delta)\sigma_{\mathcal{D}_{\text{tr}}}(\hat{\boldsymbol{\theta}}(\delta), \boldsymbol{x})$ is monotonically increasing with δ, this in turn implies

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{x}, y \sim \Pi}\Big(\hat{\varphi}(\boldsymbol{x}, y) \leq \delta\Big) = \mathbb{P}_{\boldsymbol{x}, y \sim \Pi}\left(\tilde{\beta}(\hat{\varphi}(\boldsymbol{x}, y))\sigma\left(\tilde{\boldsymbol{\theta}}\left(\hat{\varphi}(\boldsymbol{x}, y)\right), \boldsymbol{x}\right) \leq \tilde{\beta}(\delta)\sigma\left(\tilde{\boldsymbol{\theta}}\left(\delta\right), \boldsymbol{x}\right)\right) \\
=&\mathbb{P}_{\boldsymbol{x}, y \sim \Pi}\left(y - \mu_{\mathcal{D}_{\text{tr}}}(\boldsymbol{\theta}^R, \boldsymbol{x}) \leq \tilde{\beta}(\delta)\sigma\left(\tilde{\boldsymbol{\theta}}\left(\delta\right), \boldsymbol{x}\right)\right).
\end{aligned}$$

Since $\tilde{\beta}(\delta)$ and $\hat{\boldsymbol{\theta}}\left(\delta\right)$ are arbitrary, and $\hat{\beta}(\delta)$ and $\tilde{\boldsymbol{\theta}}\left(\delta\right)$ which, together with (15), implies the desired result. □

Comparison with Capone et al. (2022)

In this section, we briefly examine how our approach compares to that of Capone et al. (2022) when used to compute uniform error bounds, i.e., 100 percent credible intervals, for three different data sets. We carried out each experiment 10 times. In the following, we report the rate of uniform error bound violation and the average length of the 100 percent credible intervals. The method of Capone et al. (2022) is purely Bayesian and thus heavily dependent on the prior. The resulting credible intervals are well-calibrated, i.e., they cover most of the data. However, our approach is much better regarding sharpness. This is because Capone et al. (2022) is Bayesian and requires symmetric intervals, whereas our approach is frequentist and allows for asymmetric credible intervals. Our approach also exhibits a lower rate of uniform error bound violations than Capone et al. (2022) in most cases, which suggests that a frequentist approach is more adequate for computing uniform error bounds than a Bayesian one.

Table 2: Rate of uniform error bound violation (RUEBV) and 100% confidence interval width obtained with our approach and that of Capone et al. (2022). Lower is better for all metrics.

Data set	Metric	Ours	Capone et al. (2022)
Boston	Rate of uniform error bound violation	0.00376	**0.000172**
	Length of 100% CI	**1.2**	28.3
mpg	Rate of uniform error bound violation	**0.004**	0.0065
	Length of 100% CI	**1.7**	24.13
wine	Rate of uniform error bound violation	**0.00072**	0.00096
	Length of 100% CI	**4.7**	24.8

Bayesian Optimization

We now investigate how the proposed calibration approach can be employed in a Bayesian optimization context using two commonly used benchmark functions, the Ackley and Rosenbrock functions.

In Bayesian optimization, the goal is to find a point in input space that maximizes an unknown function $f(\cdot)$. In particular, we investigate how our calibrated GP bound performs when used as an upper confidence bound (UCB) for a GP-UCB type acquisition function. Simply put, given a data set $\mathcal{D}_t$ of size t, the GP-UCB algorithm chooses a query point by maximizing the acquisition function

$$\boldsymbol{x}_{t+1}^* = \arg\max_{\boldsymbol{x}} \mu_{\mathcal{D}_t}(\boldsymbol{\theta}^R, \boldsymbol{x}) + \beta_{\mathcal{D}_t}\sigma_{\mathcal{D}_t}(\boldsymbol{\theta}^R, \boldsymbol{x}), \tag{17}$$

where $\beta_{\mathcal{D}_t}$ is a tuning parameter that stipulates the trade-off between exploration and exploitation, and may or may not depend on the data set $\mathcal{D}_t$. It has been shown that if the unknown function $f(\cdot)$ belongs to the RKHS associated with the kernel $k(\boldsymbol{\theta}^R, \cdot, \cdot)$, and $\beta_{\mathcal{D}_t}$ is chosen sufficiently large, then the GP-UCB achieves sublinear regret (Chowdhury & Gopalan, 2017). However, both assumptions typically cannot be verified in practice, and choosing both the kernel $k(\boldsymbol{\theta}^R, \cdot, \cdot)$ and the scaling factor $\beta_{\mathcal{D}_t}$ in a principled manner remains an open problem. We propose employing the modified acquisition function

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} \mu_{\mathcal{D}_t}(\boldsymbol{\theta}^R, \boldsymbol{x}) + \beta_{\delta}\sigma_{\mathcal{D}_t}(\boldsymbol{\theta}_{\delta}, \boldsymbol{x}), \tag{18}$$

where the hyperparameters $\boldsymbol{\theta}_{\delta}$ are obtained via a calibrated model and a suitable choice of confidence parameter δ. In the experiments, we set $\beta_{\mathcal{D}_t} = 1$ and compute the calibrated hyperparameters by setting $\delta = 0.01$, meaning that we set expect only one percent of the evaluations to lie outside the confidence region. Note that even though the underlying function is fixed, it is reasonable to expect that some of the data lies outside the confidence region due to noise, and we can only expect the data to lie fully within the confidence region in the noiseless case, which we do not consider in this paper. Furthermore, we refrain from retraining the hyperparameters after each data point is collected, following the convention of other Bayesian optimization approaches (Srinivas et al., 2012; Chowdhury & Gopalan, 2017). While this does not enable us to employ the theoretical guarantees developed in Section 5, it reduces computational time significantly. We additionally compare our results to the vanilla UCB algorithm, where the hyperparameters, chosen via log-likelihood maximization, are identical for both the posterior mean and variance, and we set $\beta_{\mathcal{D}_t} = 2$.

We evaluate the results both in terms of cumulative regret and simple regret. Cumulative regret after T steps corresponds to the metric

$$R_T^{\text{cumul}} = \sum_{t=1}^{T} \left(\max_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) - f(\boldsymbol{x}_t) \right), \tag{19}$$

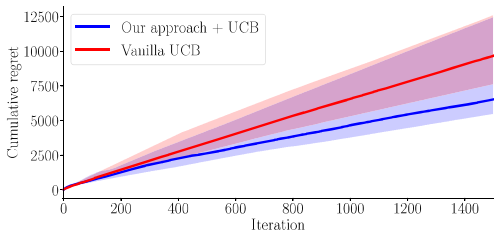
whereas simple regret is given by

$$R_T^{\text{simple}} = \max_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) - \max_{t \leq T} f(\boldsymbol{x}_t). \tag{20}$$

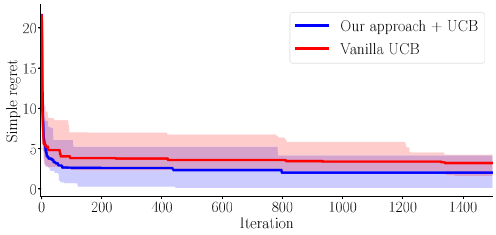
Typically, a Bayesian optimization algorithm is deemed useful if cumulative regret exhibits sublinear growth, implying that the average regret goes to zero. Simple regret, by contrast, corresponds to the best query among all past queries and is an important metric whenever evaluation costs are low (Berkenkamp et al., 2019).

In the case of the Ackley experiment, our approach typically chose lengthscales that were smaller than those computed via likelihood maximization. This results in more exhaustive exploration than vanilla UCB, which in turn means that local minima are explored more carefully before the focus of the optimization is shifted elsewhere. This results in better performance than when using vanilla UCB, both in terms of cumulative and simple regret. The results correspond to the top two figures in Figure 3.

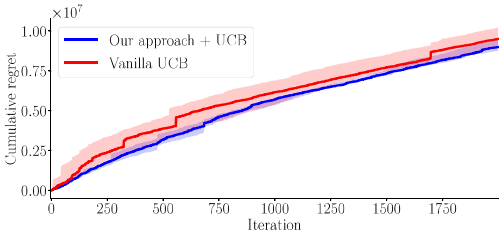
In contrast to the Ackley experiment, in the Rosenbrock experiment our approach selects lengthscales that are larger than those suggested by the likelihood maximum. Roughly speaking, this means that the confidence intervals produced by the likelihood maximum hyperparameters are too conservative, and our approach attempts to compensate for this by indicating more confidence in the posterior



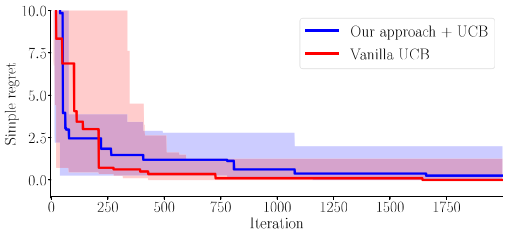
(a) Cumulative regret R_T^{cumul} of Ackley experiment.



(b) Simple regret R_T^{simple} of Ackley experiment.



(c) Cumulative regret R_T^{cumul} of Rosenbrock experiment.



(d) Simple regret R_T^{simple} of Rosenbrock experiment.

Figure 3: Regret of Ackley (top) and Rosenbrock (bottom) experiment over the number of Bayesian optimization iterations with UCB.

mean obtained with the vanilla GP. This means that local minima are explored less meticulously than with the vanilla UCB algorithm. This choice is justified by the cumulative regret obtained with our approach, as it is slightly smaller than that obtained by the vanilla UCB algorithm. However, this also results in worse simple regret than the vanilla UCB algorithm, which is intuitive, as our approach opts to explore local minima less accurately than the vanilla UCB algorithm. We also note that both algorithms converge towards the same simple regret as the number of iterations increases. The results correspond to the bottom two figures in Figure 3.