Investigation into Online Calibrated Bayesian Optimization

Teemu Turpeinen
Advisor: Marshal Sinaga
Supervisor: Samuel Kaski

Department of Computer Science
Aalto University

August 26, 2025

Background: Bayesian Optimization & Uncertainty Quantification

Bayesian Optimization (BO):

- Optimize expensive black-box:

$$x^* = \arg\max_{x \in \mathcal{X}} f(x) \qquad (1)$$

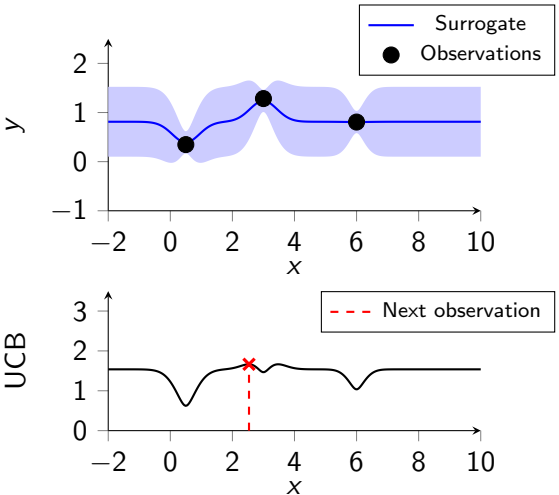
- Acquisition selects next x to evaluate

Exploration–Exploitation:

- UCB, EI, PI balance this trade-off

Why Uncertainty Matters:

- Guides decision-making
- Accurate uncertainty estimates imply a better model representation of $f(x)$



Background: Calibration and Sharpness in Bayesian Optimization

- **Uncertainty estimates** guide acquisition but can be miscalibrated
- **Calibration** ensures predicted quantiles $Q_t(p)$ match empirical coverage[1]

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{I}\{y_t \leq Q_t(p)\} \to p \text{ as } T \to \infty \tag{2}$$

- **Calibration in BO** has been achieved as a *post hoc* procedure by learning a recalibration function $R_t(p)$ using historical data[2]
  - A recalibration function simply maps target quantiles p to empirical quantiles $\hat{p}$ based on past data.
  - Achieves average calibration over past data.

[1] V. Kuleshov, N. Fenner, and S. Ermon, *Accurate uncertainties for deep learning using calibrated regression*, 2018. arXiv: 1807.00263.

[2] S. Deshpande, C. Marx, and V. Kuleshov, *Online calibrated and conformal prediction improves bayesian optimization*, 2024. arXiv: 2112.04620.

Background: Calibration and Sharpness in Bayesian Optimization

The Goal: Informative & Accurate Uncertainty

We need uncertainty estimates that are both:

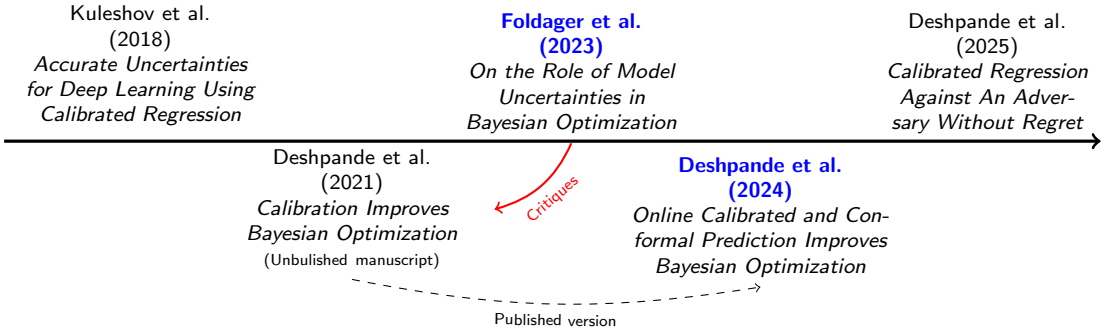
- **Calibrated (Accurate):** Are our 90% credible intervals correct 90% of the time?
- **Sharp (Informative):** How narrow are the credible intervals?

Optimal forecasts should thus be maximally *sharp* subject to calibration (2):

$$\min_{\hat{Q}_p} \mathbb{E}_x \left[|\hat{Q}_p(x) - \hat{Q}_{1-p}(x)| \right]$$

While many approaches exist to balance this trade-off, our work focuses on a subfield. We will now give a brief snapshot of Online Calibrated Bayesian Optimization.

Snapshot of Online Calibrated Bayesian Optimization



Other notable works

- Kuleshov et al. (2022) - *Calibrated and Sharp Uncertainties in Deep Learning via Density Estimation*
- Capone et al. (2023) - *Sharp Calibrated Gaussian Processes*
- Stanton et al. (2023) - *Bayesian Optimization with Conformal Prediction Sets*

Critique and proposed method

Critiques and key challenges
(Foldager et al., 2023)

- Calibration is not necessarily correlated with BO performance
- Calibration is not reliable for small non-i.i.d. data samples
- Recalibration can in some cases lead to more miscalibration

Online Recalibration
(Deshpande et al., 2024)

To address this, the online method recalibrates sequentially by minimizing a loss function $\bar{l}_{sp}(y_s, q)$ against miscalibration, which provides asymptotic guarantees for the level of calibration.

$$R_t(p) = \inf \arg\min_q \left[\frac{1}{2\eta^2} + \sum_{s=1}^{t-1} \bar{l}_{sp}(y_s, q) \right]$$

- $\frac{1}{2\eta^2}$ is a regularization term controlling update size through η

Key research points:

1. How are calibration, sharpness and proper scores fundamentally connected in BO performance?
2. Is Deshpande's new online method fundamentally different from the older method that the original critique was based on?
3. Ultimately, what should we focus on to achieve better BO performance?

Practical questions

In designing the experiments and reading source code, several methodological ambiguities were revealed:

- **Moment vs. Quantile Recalibration:** Foldager et al. recalibrate the entire predictive distribution (approximating it as Gaussian via moment estimation), whereas other methods directly adjust the quantiles of the posterior, effectively tuning its variance.
- **The Undocumented η Parameter:** The learning rate η in the online method is critical, but its values or tuning process are unspecified.
- **Test Set Construction:** The methods assess calibration differently: Foldager et al. use a global, independent test set, while the online approach uses each query as a test point before updating the surrogate model.

Expanded Objective

Therefore, our final goal was to design experiments that could simultaneously resolve these practical questions while also shedding light on our initial theoretical questions.

Metrics

Simple Regret (r_t)
Difference between current best and optimal decision. *Lower is better.*

$$r_t = f(x^*) - f(x_t^{\text{best}})$$

Cumulative Regret (R_T)
Measures the total opportunity cost over time. *Lower is better.*

$$R_T = \sum_{t=1}^{T} \left(f(x^*) - f(x_t) \right)$$

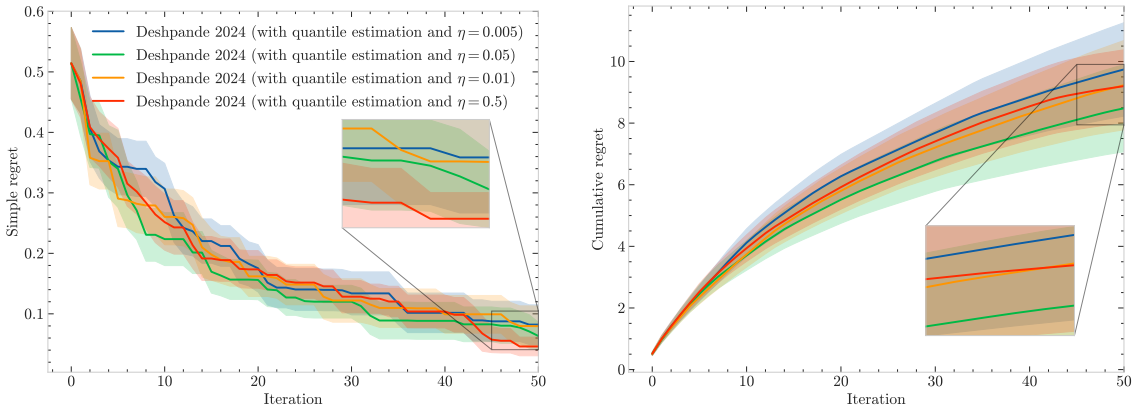
Expected Calibration Error (ECE)
Measures how well predicted probabilities p match empirical frequencies $\hat{p}$. *Lower is better.*

$$\text{ECE} = \frac{1}{P} \sum_{j=1}^{P} (\hat{p}_j - p_j)^2$$

Continuous Ranked Probability Score (CRPS)
Measures the integrated squared difference between the predicted CDF $F(y)$ and the empirical CDF of the single observed outcome x, which is represented by the Heaviside step function $H(y - x)$. *Lower is better.*

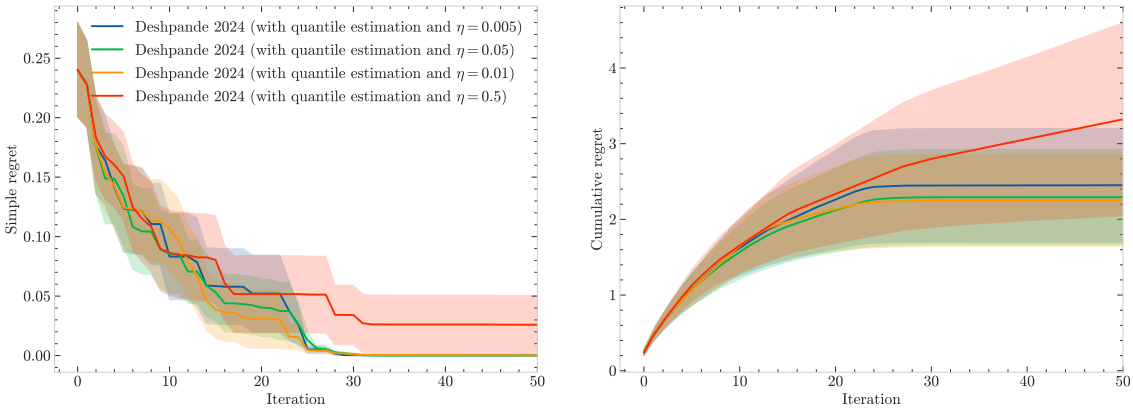
$$\text{CRPS}(F, x) = \int_{-\infty}^{\infty} \left(F(y) - H(y - x) \right)^2 dy$$

Sensitivity to Learning Rate η (with noise)



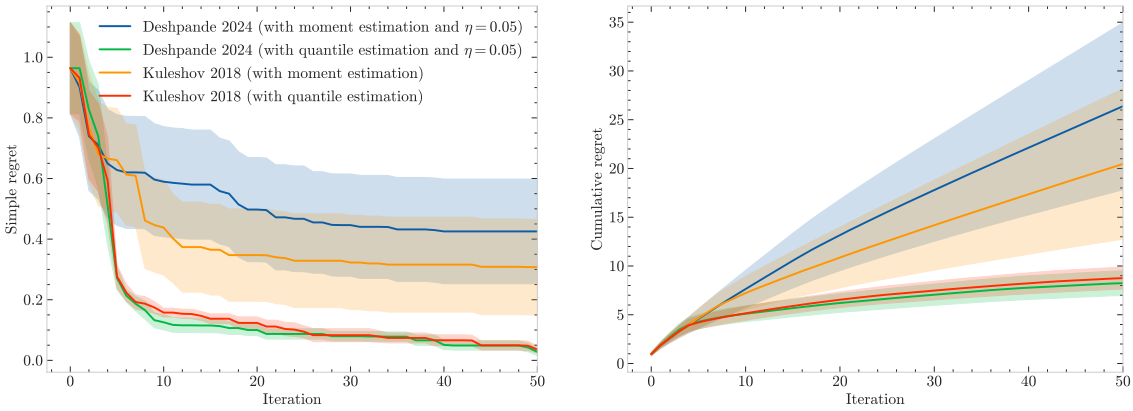
Figures 1-2: Effect of learning rate η evaluated on simple and cumulative regret with the Six-hump-camel Function (2D) and UCB.

Sensitivity to Learning Rate η (without noise)



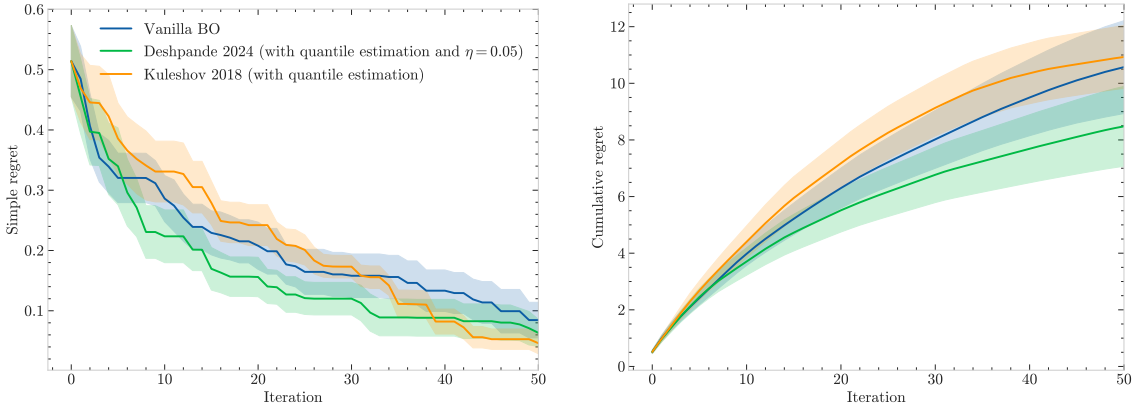
Figures 1-2: Effect of learning rate η evaluated on simple and cumulative regret with the Six-hump-camel (2D) benchmark and UCB acquisition function.

Quantile vs. Moment Estimation



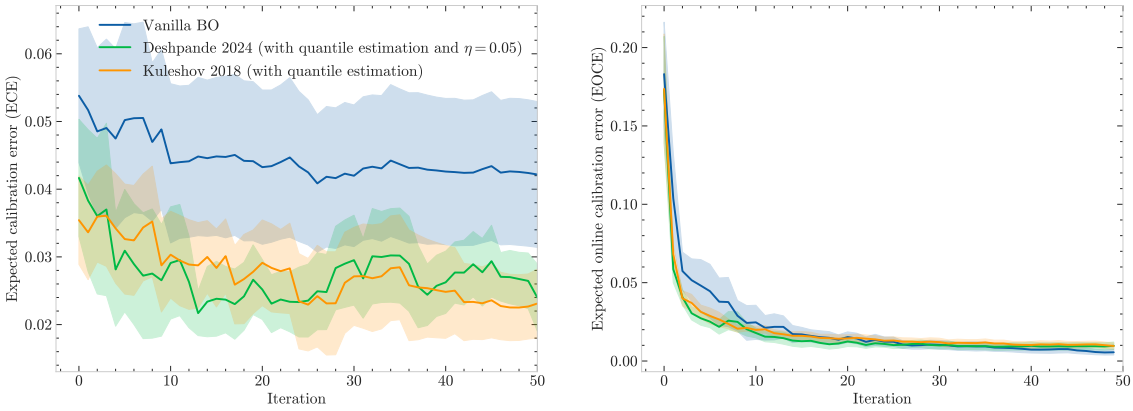
Figures 3-4: Moment vs quantile estimation evaluated on simple and cumulative regret with the Forrester (1D) benchmark and UCB acquisition function.

Performance differences



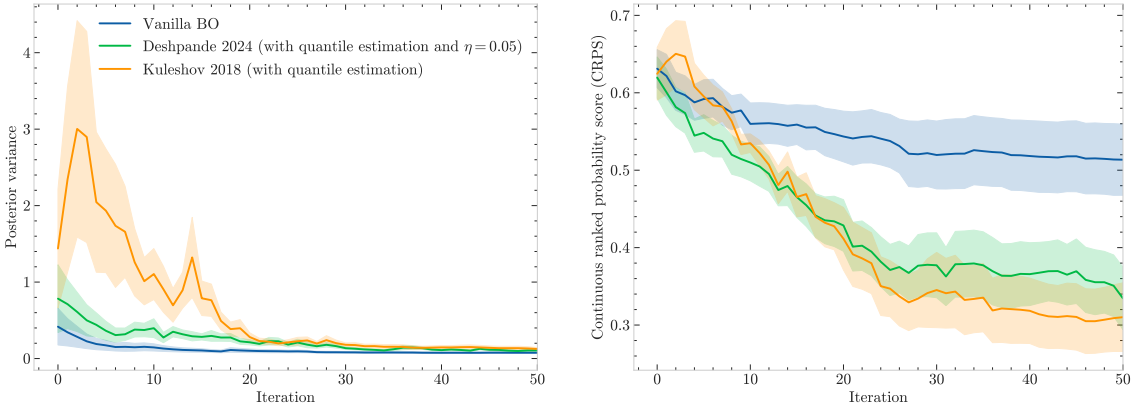
Figures 5-6: Simple and cumulative regret evaluated with the Six-hump-camel (2D) benchmark and UCB acquisition function.

Calibration



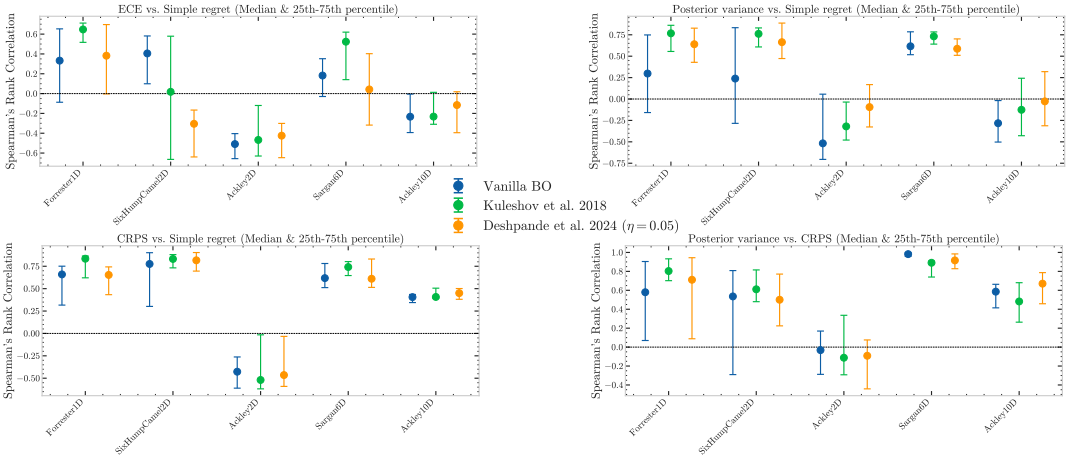
Figures 7-8: Expected Calibration Error evaluated with a global (ECE) and online (EOCE) test sets. Benchmark Six-hump-camel (2D) benchmark and UCB acquisition function.

Sharpness and CRPS



Figures 9-10: Posterior variance and continuous ranked probability score (CRPS) on Six-hump-camel (2D) benchmark and UCB acquisition function.

Correlations



Finding

ECE and sharpness alone are not as strongly correlated with regret as CRPS.

Conclusions

Answers to Practical Questions

- Quantile recalibration is more stable than the moment estimation approach.
- The online method's performance is highly sensitive to the learning rate η. An effective but still stable learning rate was $\eta = 0.05$.
- The choice of test set (global vs. online) significantly impacts the interpratation of calibration performance.

Answers to Research Questions

- Low calibration error or sharpness is not a guarantee of low simple regret, although for some benchmarks there is a clear positive correlation.
- The critiques by Foldager et al. are still valid although the performance of recalibrated BO is better with the online approach.

$\rightarrow$ BO performance should not be solely assessed through calibration or sharpness metrics but there should be a balance given by e.g. proper scores like CRPS.

´

What next?

Experiments

- Assess performance using proper scores to better understand the balance between calibration and sharpness.
- Investigate theoretical claims in a controlled setting with bounded functions (e.g., GP sample paths), where regret-bound assumptions hold.

Theory

- Deriving a cumulative regret bound that is an explicit function of both sharpness and calibration.[3][4]
- Calibrated mean variance reduction

Surrogate model

- Interpretable calibrated surrogate model e.g. a Sharp calibrated GP.[4]

---

[3] N. Srinivas, A. Krause, S. M. Kakade, *et al.*, "Information-theoretic regret bounds for gaussian process optimization in the bandit setting," *IEEE Transactions on Information Theory*, no. 5, 3250–3265, May 2012. DOI: 10.1109/tit.2011.2182033.

[4] A. Capone, G. Pleiss, and S. Hirche, *Sharp calibrated gaussian processes*, 2023. arXiv: 2302.11961.

References

[1] V. Kuleshov, N. Fenner, and S. Ermon, *Accurate uncertainties for deep learning using calibrated regression*, 2018. arXiv: 1807.00263.

[2] S. Deshpande, C. Marx, and V. Kuleshov, *Online calibrated and conformal prediction improves bayesian optimization*, 2024. arXiv: 2112.04620.

[3] N. Srinivas, A. Krause, S. M. Kakade, *et al.*, "Information-theoretic regret bounds for gaussian process optimization in the bandit setting," *IEEE Transactions on Information Theory*, no. 5, 3250–3265, May 2012. DOI: 10.1109/tit.2011.2182033.

[4] A. Capone, G. Pleiss, and S. Hirche, *Sharp calibrated gaussian processes*, 2023. arXiv: 2302.11961.