A Deep Dive Into BO Sensitivity and PROBO

Julian Rodemann, Thomas Augustin

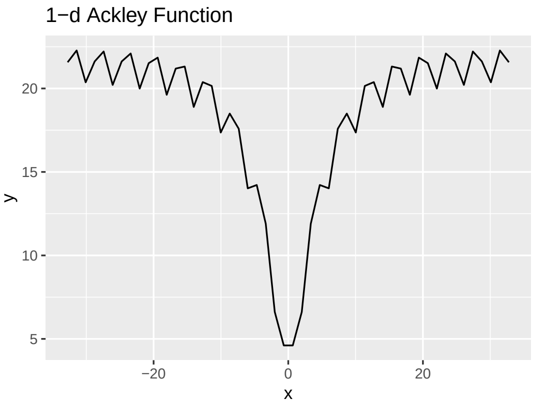
Young Statisticians Lecture Series (YSLS)
IBS-DR Early Career Working Group

May 4, 2022

Agenda

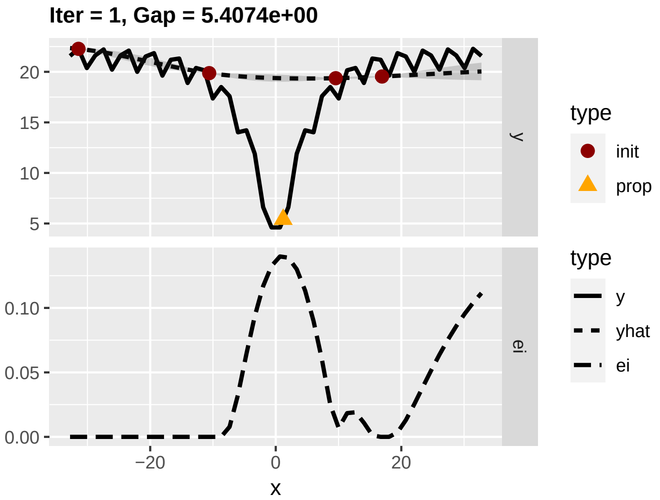
Agenda

Bayesian Optimization

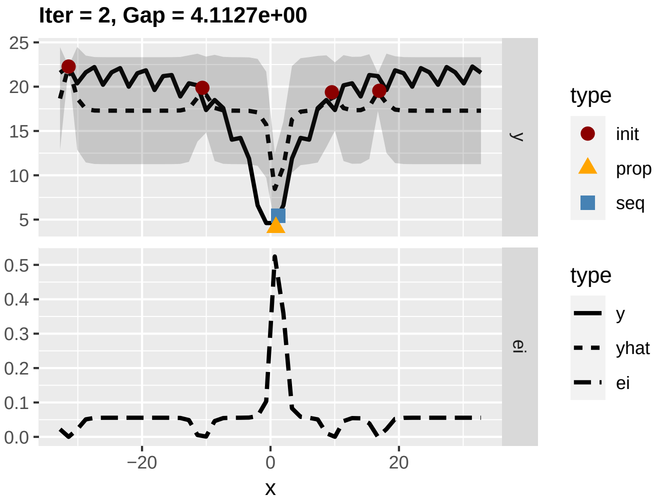


Note: If not otherwise stated, all figures are based on own computations using `ggplot2` [Wickham, 2016], `smoof` [Bossek, 2017] and `mlr(3)MBO` [Bischl et al., 2017]

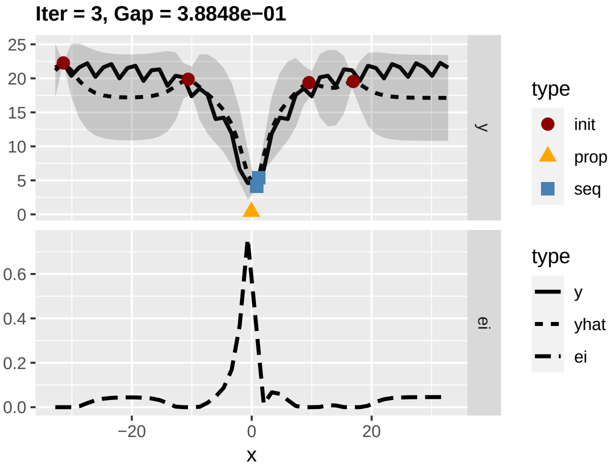
Bayesian Optimization



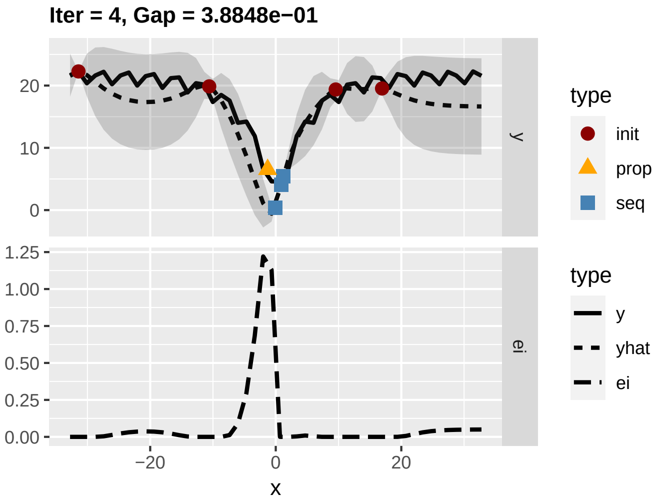
Bayesian Optimization



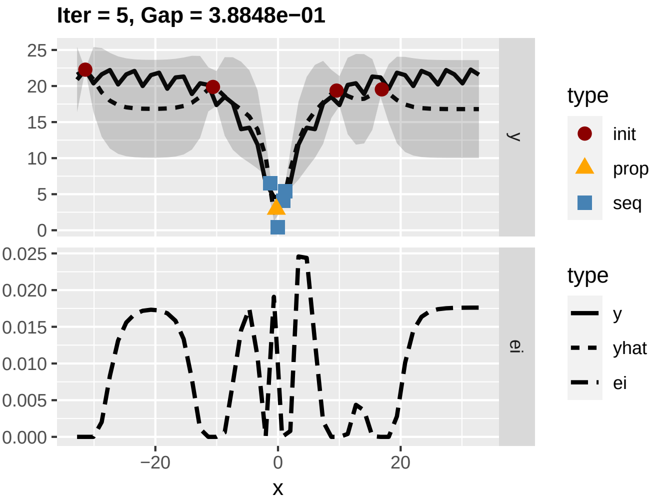
Bayesian Optimization



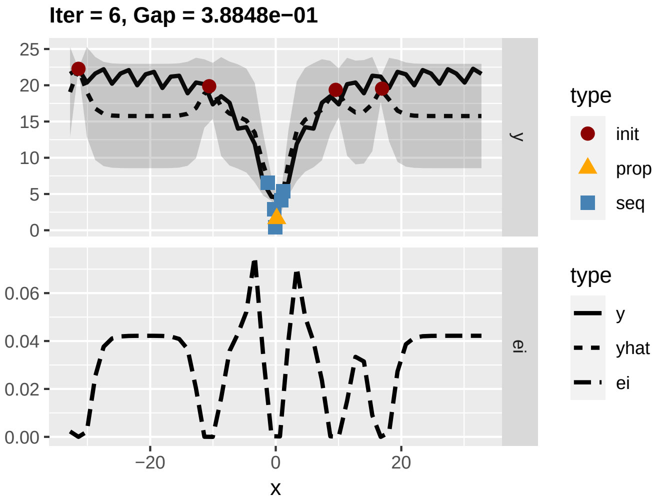
Bayesian Optimization



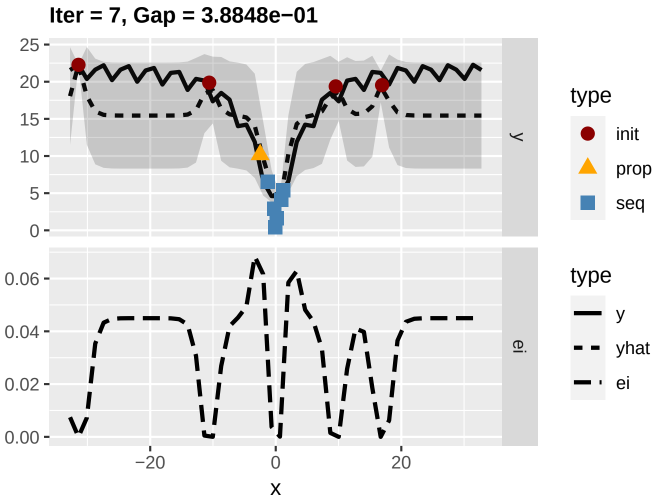
Bayesian Optimization



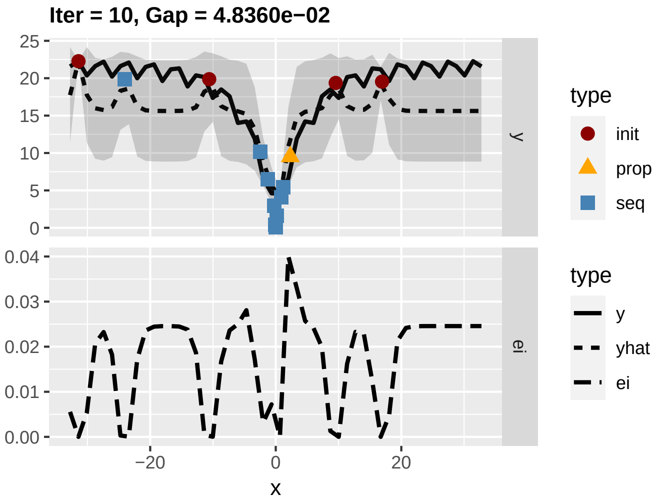
Bayesian Optimization



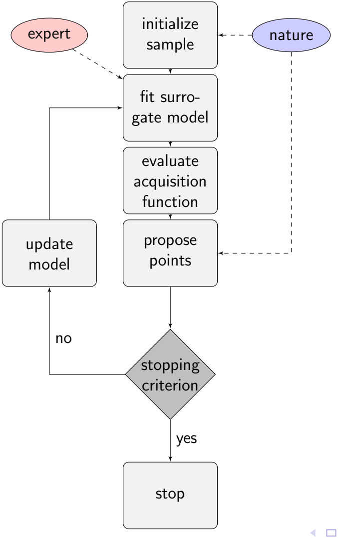
Bayesian Optimization



Bayesian Optimization



Bayesian Optimization



BO: Some Applications

- Hyperparamter-tuning, e.g. AlphaGo [Chen et al., 2018]
- Engineering [Frazier and Wang, 2016] [Jones et al., 1998]
- Cognitive science [Shi et al., 2013]
- Climate modeling [Abbas et al., 2014]
- Drug discovery [Pyzer-Knapp, 2018]
  - "prioritizing molecules within the discovery process"
- Or more recently COVID-19 detection [Awal et al., 2021]

Agenda

Gaussian Processes

Definition (Gaussian Process Regression)

A function $f(\boldsymbol{x})$ is generated by a Gaussian process $\mathcal{GP}\left(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right)$ if for any finite set of data points $\{x_1, \ldots, x_n\}$, the associated vector of function values $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))$ has a multivariate Gaussian distribution: $\boldsymbol{f} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$.

Note: For a comprehensive introduction to Gaussian process regression see [Rasmussen, 2003].

Gaussian Processes - Intuition

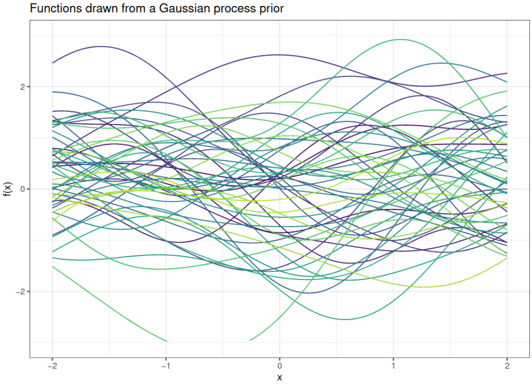


Image credits: [Moosbauer and Bischl, 2019]

Gaussian Processes - Intuition

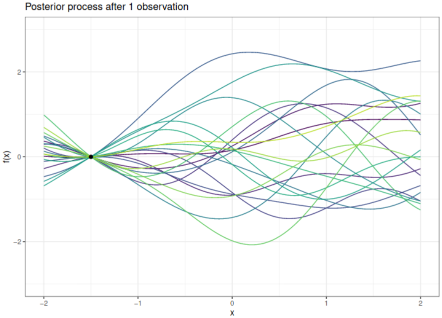


Image credits: [Moosbauer and Bischl, 2019]

Gaussian Processes - Intuition

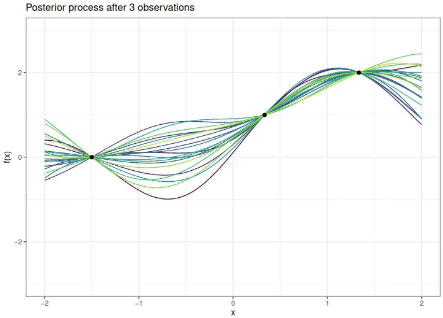
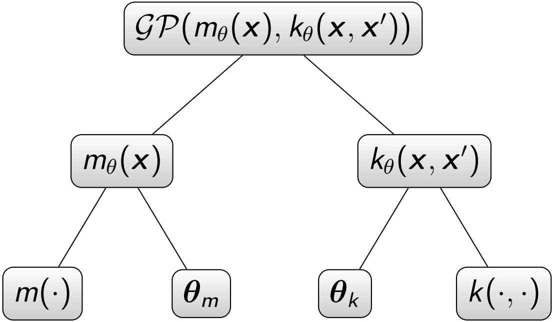


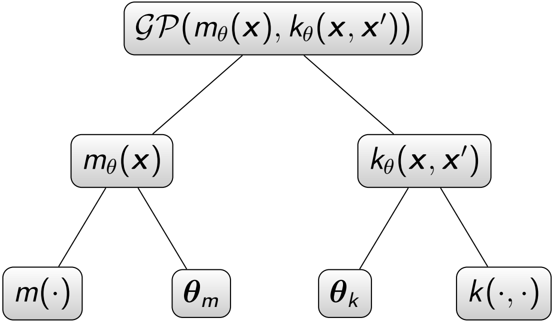
Image credits: [Moosbauer and Bischl, 2019]

Gaussian Processes – Prior Components



How to specify $m(\cdot)$, $\boldsymbol{\theta}_m$, $\boldsymbol{\theta}_k$ and $k(\cdot, \cdot)$
in absence of prior knowledge?

Gaussian Processes – Prior Components



And: Do they even matter?

Agenda

Agenda

Setup

- We randomly select 50 synthetic test functions from the `R` package `smoof` [Bossek, 2017], stratified across the covariate space dimensions $1, 2, 3, 4$ and 7.
- For each of them, a sensitivity analysis is conducted with regard to each of the four prior components.
  - 5 functional forms
  - 5 mean and kernel parameter specifications (relative deviation from global mean)
  - we control for interaction effects
- The initial design of size $n_{init} = 10$ is randomly sampled anew for each of the $R = 40$ BO repetitions with $T = 20$ iterations each.

Mean Optimization Path

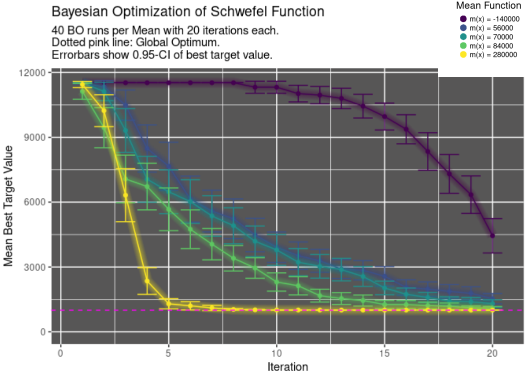
Definition (Mean Optimization Path)

Given R repetitions of Bayesian optimization applied on a test function $\Psi(\boldsymbol{x})$ with T iterations each, let $\Psi(\boldsymbol{x}^*)_{r,t}$ be the best incumbent target value at iteration $t \in \{1, ..., T\}$ from repetition $r \in \{1, ..., R\}$. The elements

$$MOP_t = \frac{1}{R} \sum_{r=1}^{R} \Psi(\boldsymbol{x}^*)_{r,t}$$

shall then constitute the T-dimensional vector MOP, which we call *mean optimization path (MOP)* henceforth.

Example: MOPs for BO on Schwefel Function



Accumulated Difference of MOPs

Definition (Accumulated Difference of MOPs)

Consider an experiment comparing S different prior specifications on a test function with R repetitions per specification and T iterations per repetition. Let the results be stored in a $T \times S$-matrix of mean optimization paths for iterations $t \in \{1, ..., T\}$ and prior specification $s \in \{1, ..., S\}$ (e.g. constant, linear, quadratic etc. trend as mean functional form) with entries $MOP_{t,s} = \frac{1}{R} \sum_{r=1}^{R} \Psi(\boldsymbol{x}^*)_{r,t,s}$. The *accumulated difference (AD)* for this experiment shall then be:

$$AD = \sum_{t=1}^{T} \left(\max_s MOP_{t,s} - \min_s MOP_{t,s} \right).$$

Agenda

Results

Mean functional form	Kernel functional form	Mean parameters	Kernel parameters
42.49	68.20	77.91	11.40

Table: Sum of relative ADs of all 50 MOPs per prior specification.

Agenda

Agenda

Prior near-ignorance models

- Idea: Use set of $\boldsymbol{\theta}_m$ instead of precise $\boldsymbol{\theta}_m$. Fully specify the other components.
- [Mangili, 2015] proposes imprecise Gaussian processes

$$\left\{\mathcal{GP}\left(Mh, k_{\boldsymbol{\theta}}(x, x') + \frac{1+M}{c}\right) : h = \pm 1, M \geq 0\right\},$$

given a base kernel $k_{\boldsymbol{\theta}}(x, x')$ and a degree of imprecision $c > 0$.

$\rightarrow$ results in a set of posteriors whose upper and lower mean estimates $\underline{\hat{\mu}}(x)_c$, $\overline{\hat{\mu}}(x)_c$ can be derived

Note: See [Benavoli and Zaffalon, 2015] for an introduction to prior near-ignorance models.

Upper and lower mean estimates

In order to derive upper and lower bounds for the mean estimate, let $k_\theta(x, x')$ be a kernel function as defined in [Rasmussen, 2003]. The finitely positive semi-definite matrix $\boldsymbol{K}_n$ is then formed by applying $k_\theta(x, x')$ on the training data vector $\mathbf{x}$:

$$\boldsymbol{K}_n = [k_\theta(x_i, x_j')]_{ij}. \tag{1}$$

Let x be a scalar input of test data, whose $f(x)$ is to be predicted. Then $\boldsymbol{k}_x = [k_{\boldsymbol{\theta}}(x, x_1), ..., k_{\boldsymbol{\theta}}(x, x_n)]^T$ is the vector of covariances between x and the training data. Furthermore, name the training target vector $\boldsymbol{y}$ and define $\boldsymbol{s}_k = \boldsymbol{K}_n^{-1} 1_n$ as well as $\boldsymbol{S}_k = 1_n^T \boldsymbol{K}_n^{-1} 1_n$.

Upper and lower mean estimates

Then [Mangili, 2015] shows that if $|\frac{\boldsymbol{s}_k \boldsymbol{y}}{\boldsymbol{S}_k}| \leq 1 + \frac{c}{\boldsymbol{S}_k}$:

$$\overline{\hat{\mu}}(x) = \boldsymbol{k}_x^T \boldsymbol{K}_n^{-1} \boldsymbol{y} + (1 - \boldsymbol{k}_x^T \boldsymbol{s}_k) \frac{\boldsymbol{s}_k^T}{\boldsymbol{S}_k} \boldsymbol{y} + c \frac{|1 - \boldsymbol{k}_x^T \boldsymbol{s}_k|}{\boldsymbol{S}_k} \quad (2)$$

$$\underline{\hat{\mu}}(x) = \boldsymbol{k}_x^T \boldsymbol{K}_n^{-1} \boldsymbol{y} + (1 - \boldsymbol{k}_x^T \boldsymbol{s}_k) \frac{\boldsymbol{s}_k^T}{\boldsymbol{S}_k} \boldsymbol{y} - c \frac{|1 - \boldsymbol{k}_x^T \boldsymbol{s}_k|}{\boldsymbol{S}_k} \quad (3)$$

Upper and lower mean estimates

If $\left|\frac{\mathbf{s}_k \mathbf{y}}{\mathbf{S}_k}\right| > 1 + \frac{c}{\mathbf{S}_k}$:

$$\overline{\hat{\mu}}(x) = \mathbf{k}_x^T \mathbf{K}_n^{-1} \mathbf{y} + (1 - \mathbf{k}_x^T \mathbf{s}_k) \frac{\mathbf{s}_k^T}{\mathbf{S}_k} \mathbf{y} + c \frac{1 - \mathbf{k}_x^T \mathbf{s}_k}{\mathbf{S}_k} \tag{4}$$

$$\underline{\hat{\mu}}(x) = \mathbf{k}_x^T \mathbf{K}_n^{-1} \mathbf{y} + (1 - \mathbf{k}_x^T \mathbf{s}_k) \frac{\mathbf{s}_k^T \mathbf{y}}{c + \mathbf{S}_k} \tag{5}$$

Agenda

Hedging (1)

- deploy several $\underline{\mu}(x)_c$, $\overline{\mu}(x)_c$ for varying c as SMs in parallel
- return $2S + 1$ optima for S imprecise surrogate models and the precise model
- $2S$ additionally proposed optima hedge against prior misspecification
- provides "out-of-the-bag" sensitivity analysis

  $\rightarrow$ stopping criterion?

Agenda

Batches (2)

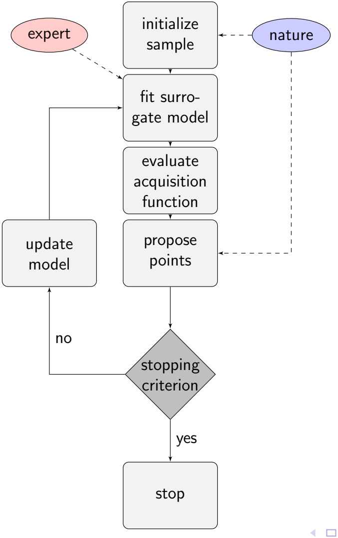
- define initial budget $K+1$ of Cores with $S=\frac{K}{2}+1$ (I)GP models (as in 1.)
- distribute budget B of total evaluations among M batches and respective number of Cores $C \in \mathbb{N}^M$ with $C=(K+1, \left\lfloor \frac{K+1}{2} \right\rfloor, \left\lfloor \frac{K+1}{4} \right\rfloor, \dots)$
- after each $m \in M$ dismiss worst $\frac{K}{2}$ models

Agenda

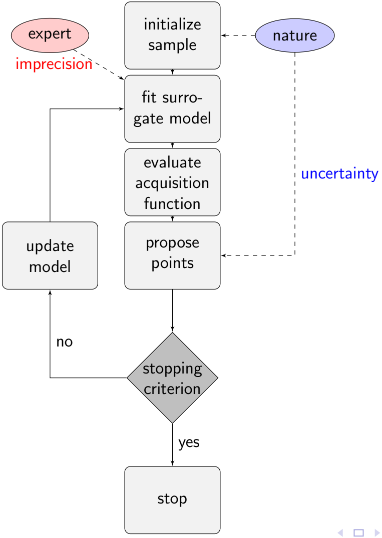
Generalized Lower Confidence Bound (GLCB)

- $LCB(x) = -\widehat{\mu}(x) + \tau \cdot \underbrace{\sqrt{\widehat{\text{Var}}(\mu(x))}}_{\text{"classical" uncertainty}}$
- $GLCB(x) = -\widehat{\mu}(x) + \tau \cdot \underbrace{\sqrt{\widehat{\text{Var}}(\mu(x))}}_{\text{"classical" uncertainty}} + \rho \cdot \underbrace{(\overline{\mu}(x)_c - \underline{\mu}(x)_c)}_{\text{prior-induced imprecision}}$
  - τ is the degree of risk-aversion
  - ρ is the degree of ambiguity-aversion

Bayesian Optimization



Bayesian Optimization



Generalized Lower Confidence Bound (GLCB)

Notably, $\overline{\hat{\mu}}(\boldsymbol{x}) - \underline{\hat{\mu}}(\boldsymbol{x})$ simplifies to an expression only dependent on predictive kernels $\boldsymbol{k}_x = [k_{\boldsymbol{\theta}}(x, x_1), \ldots, k_{\boldsymbol{\theta}}(x, x_n)]^T$, the base kernel matrix $\boldsymbol{K}_n$ (from training) and the degree of imprecision c. If $|\frac{\boldsymbol{s}_k \boldsymbol{y}}{\boldsymbol{S}_k}| > 1 + \frac{c}{\boldsymbol{S}_k}$:

$$\overline{\hat{\mu}}(x) - \underline{\hat{\mu}}(x) = (1 - \boldsymbol{k}_x^T \boldsymbol{s}_k)(\frac{\boldsymbol{s}_k^T}{\boldsymbol{S}_k}\boldsymbol{y} + \frac{c}{\boldsymbol{S}_k} - \frac{\boldsymbol{s}_k^T \boldsymbol{y}}{c + \boldsymbol{S}_k}) \qquad (6)$$

Generalized Lower Confidence Bound (GLCB)

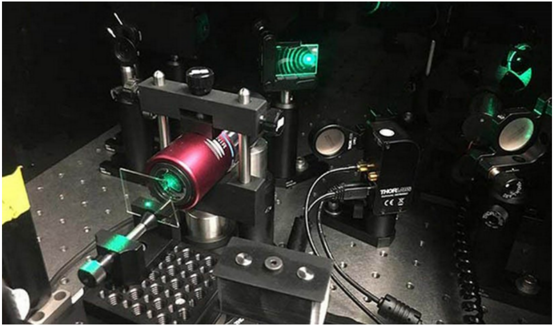
For sufficiently high c, the model imprecision $\overline{\hat{\mu}}(\boldsymbol{x}) - \underline{\hat{\mu}}(\boldsymbol{x})$ even simplifies further:

$$\overline{\hat{\mu}}(x) - \underline{\hat{\mu}}(x) = 2c\frac{|1 - \boldsymbol{k}_x^T \boldsymbol{s}_k|}{\boldsymbol{S}_k} \tag{7}$$

In this case, GLCB's hyperparameters ρ and c collapse to one.

Agenda

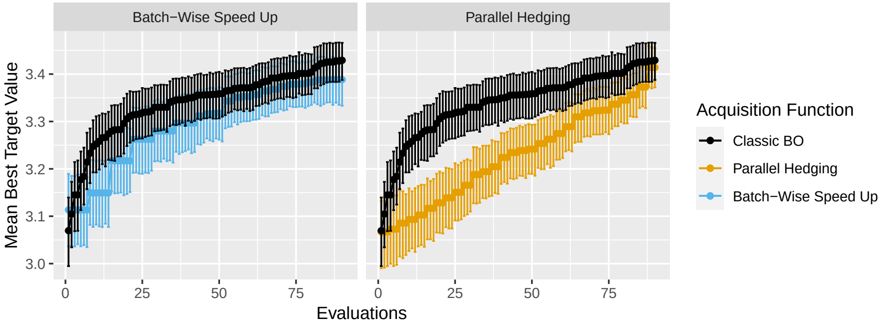
Application in Material Science



Experimental set-up of graphene production: "The preparation of a sample to be irradiated requires about **one week**." [Kotthoff, 2019]

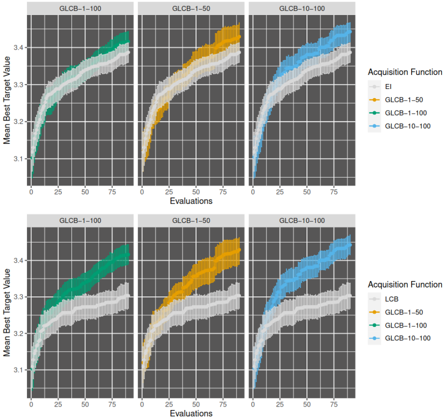
Image credits: Lars Kotthoff, University of Wyoming

Results – Hedge (1) and Batch (2)



Benchmarking results from BO on Graphene quality function. Data source: [Wahab et al., 2020].

Results – GLCB (3)



BO with GLCB on Graphene function. GLCB-1-50 means GLCB with $\rho = 1$, $c = 50$. Data source: [Wahab et al., 2020].

Agenda

Discussion

- Limitations
  - robust only with regard to possible misspecification of the mean function parameter given a constant trend
  - how to specify c?
- Venues for future work
  - locally
    - Can we ensure $|\frac{\boldsymbol{s}_k \boldsymbol{y}}{\boldsymbol{S}_k}| \leq 1 + \frac{c}{\boldsymbol{S}_k}$ such that hyperparameters c and ρ collapse to one?
  - globally
    - Imprecise probabilities offer vivid framework to represent ignorance in surrogate-assisted derivative-free optimization

Discussion

- Thanks a lot for your attention!
- Feel free to try out PROBO yourself: `https://github.com/rodemann/gp-imprecision-in-bo`
- We are looking forward to your feedback and comments of any kind!

PROBO: Literature

- Rodemann, J.: *Robust Generalizations of Stochastic Derivative-Free Optimization.* Master's thesis, LMU Munich (2021) [1]
- Rodemann, J., Augustin, T.: *Accounting for Gaussian Process Imprecision in Bayesian Optimization.* In: Honda, K., Entani, T., Ubukata, S., Huynh, V.N., Inuiguchi, M. (eds.) IUKM. Springer Lecture Notes in Computer Science (LNCS). pp. 92–104. Springer, Cham (2022)

[1] https://epub.ub.uni-muenchen.de/77441/1/MA_Rodemann.pdf

Agenda

Literature I

Abbas, M., Ilin, A., Solonen, A., Hakkarainen, J., Oja, E., and Järvinen, H. (2014).
Bayesian optimization for tuning chaotic systems.
Nonlinear Processes in Geophysics Discussions, 1(2):1283–1312.

Awal, M. A., Masud, M., Hossain, M. S., Bulbul, A. A., Mahmud, S. M. H., and Bairagi, A. K. (2021).
A novel Bayesian optimization-based machine learning framework for COVID-19 detection from inpatient facility data.
IEEE Access, 9:10263–10281.

Literature II

Benavoli, A. and Zaffalon, M. (2015).
Prior near ignorance for inferences in the k-parameter exponential family.
Statistics, 49(5):1104–1140.

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., and Lang, M. (2017).
mlrmbo: A modular framework for model-based optimization of expensive black-box functions.
arXiv preprint arXiv:1703.03373.

Literature III

Bossek, J. (2017).
smoof: Single- and multi-objective optimization test functions.
The R Journal.

Chen, Y., Huang, A., Wang, Z., Antonoglou, I., Schrittwieser, J., Silver, D., and de Freitas, N. (2018).
Bayesian optimization in alphago.
arXiv preprint arXiv:1812.06855.

Frazier, P. I. and Wang, J. (2016).
Bayesian optimization for materials design.
In *Information Science for Materials Discovery and Design*, pages 45–75. Springer.

Literature IV

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.

Kotthoff, L. (2019).
Ai for materials science: Tuning laser-induced graphene production and beyond.

Literature V

Mangili, F. (2015).
A prior near-ignorance Gaussian process model for nonparametric regression.
In *ISIPTA '15: Proceedings of the 9th International Symposium on Imprecise Probability: Theories and Applications*, pages 187–196.

Moosbauer, J. and Bischl, B. (2019).
Fortgeschrittene computerintensive Methoden - Lecture slides (LMU, Summer term 2019).

Pyzer-Knapp, E. O. (2018).
Bayesian optimization for accelerated drug discovery.
IBM Journal of Research and Development, 62(6):2–1.

Literature VI

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

Shi, Z., Church, R. M., and Meck, W. H. (2013).
Bayesian optimization of time perception.
Trends in cognitive sciences, 17(11):556–564.

Literature VII

Wahab, H., Jain, V., Tyrrell, A. S., Seas, M. A., Kotthoff, L., and Johnson, P. A. (2020).
Machine-learning-assisted fabrication: Bayesian optimization of laser-induced graphene patterning using in-situ raman analysis.
Carbon, 167:609–619.

Wickham, H. (2016).
ggplot2: Elegant Graphics for Data Analysis.
Springer-Verlag New York.

Agenda