Julian Rodemann[1], Thomas Augustin[1]

Towards Prior-Mean Robust Bayesian Optimization

Young Statisticians Session (YSS)

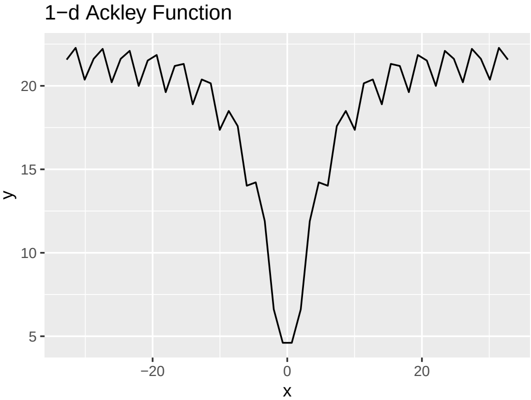
DAGStat 2022

March 30, 2022, Hamburg

[1]Department of Statistics, LMU Munich

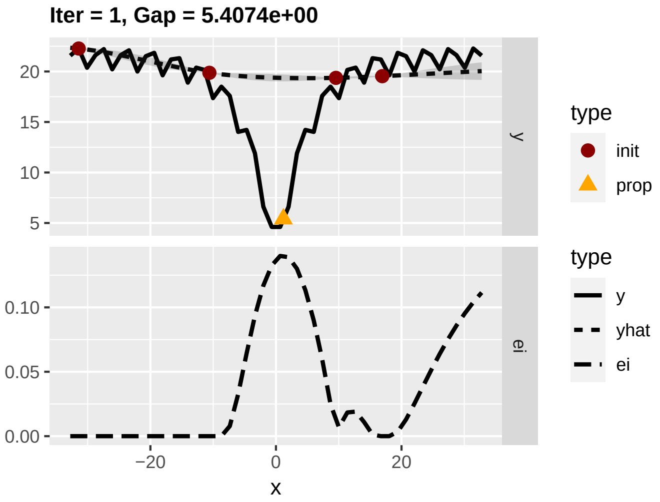
Agenda

Agenda

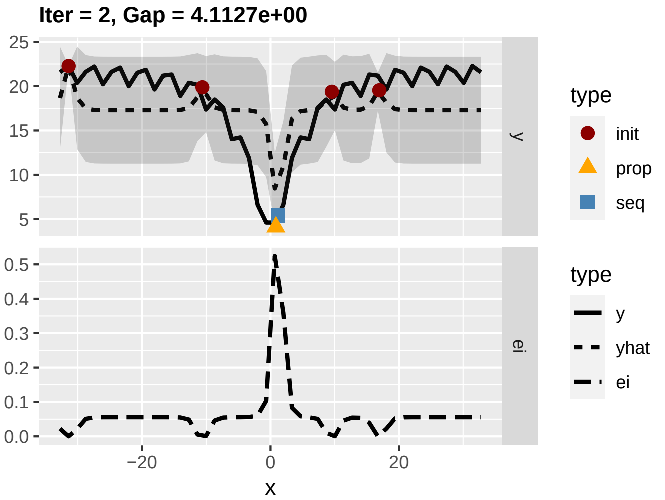


.

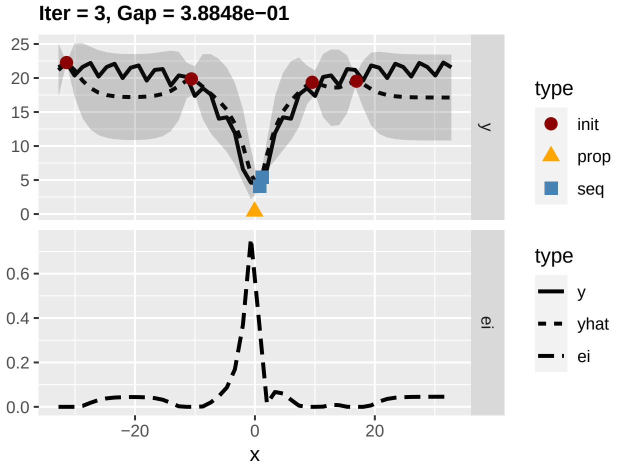
Note: If not otherwise stated, all figures are based on own computations using `ggplot2` [Wickham, 2016], `smoof` [Bossek, 2017] and `mlr(3)MBO` [Bischl et al., 2017]



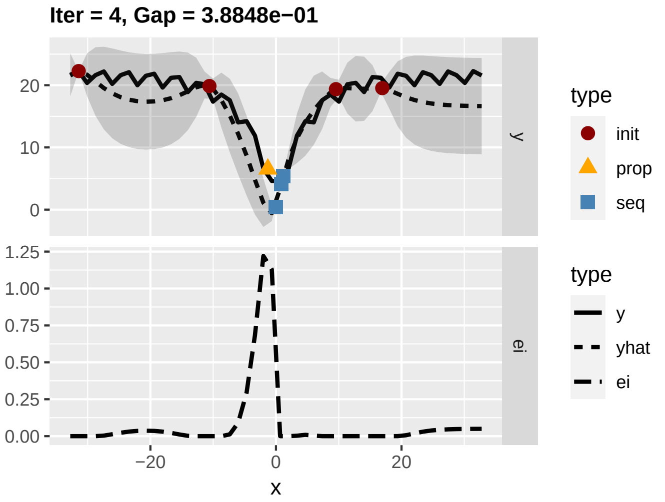
Iteration 1: Surrogate Model (top) and Acquisition Function (bottom)


Iter = 2, Gap = 4.1127e+00
25
20
15
10
5
0.5
0.4
0.3
0.2
0.1
0.0
−20
0
20
x
y
ei
type
init
prop
seq
type
y
yhat
ei

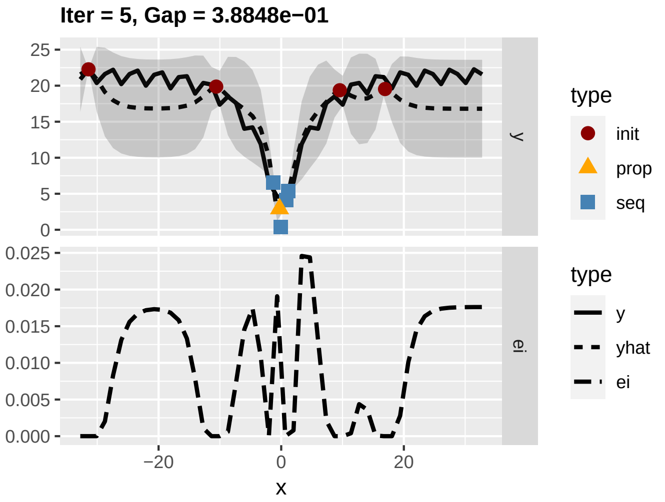
Iteration 2


Iter = 3, Gap = 3.8848e−01
25
20
15
10
5
0
y
0.6
0.4
0.2
0.0
ei
−20
0
20
x
type
init
prop
seq
type
y
yhat
ei

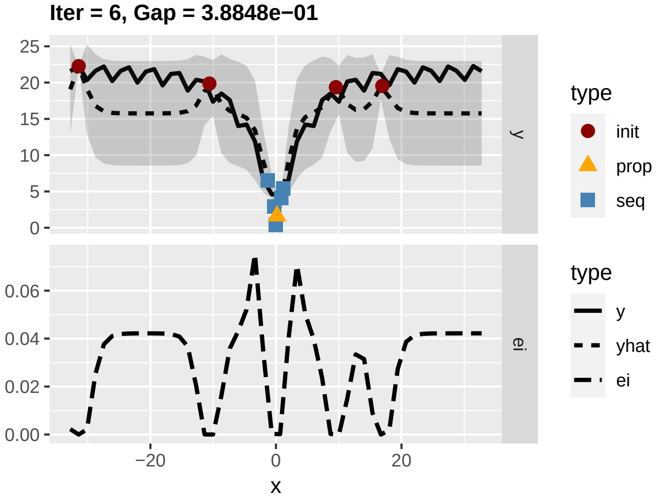
Iteration 3


Iter = 4, Gap = 3.8848e−01
20
10
0
y
1.25
1.00
0.75
0.50
0.25
0.00
ei
−20
0
20
x
type
init
prop
seq
type
y
yhat
ei

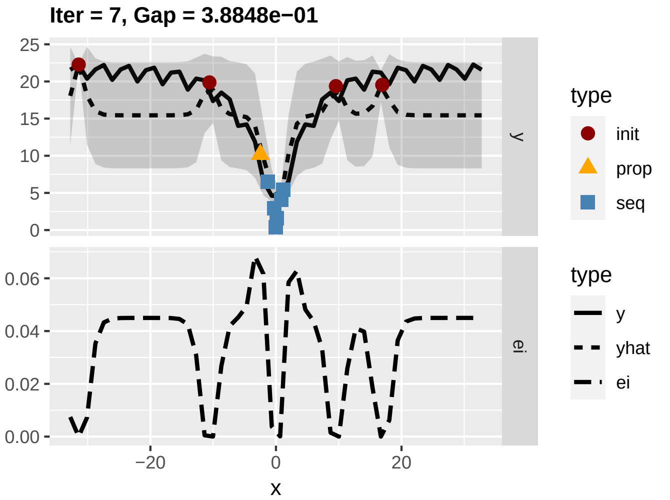
Iteration 4


Iter = 5, Gap = 3.8848e−01
25
20
15
10
5
0
0.025
0.020
0.015
0.010
0.005
0.000
−20
0
20
x
y
ei
type
init
prop
seq
type
y
yhat
ei

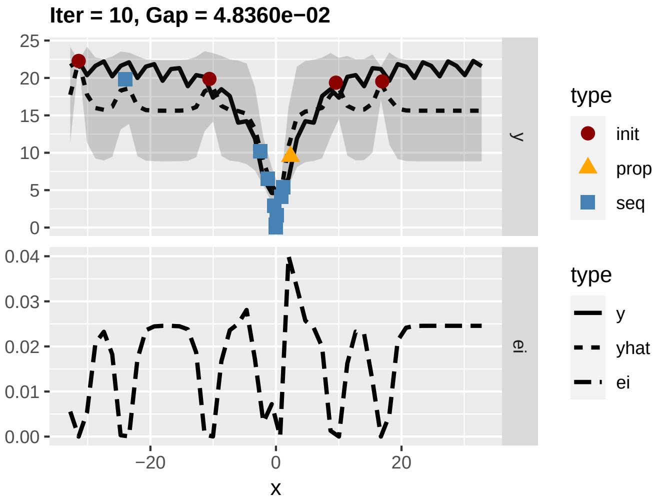
Iteration 5


Iter = 6, Gap = 3.8848e−01
25
20
15
10
5
0
y
0.06
0.04
0.02
0.00
ei
−20
0
20
x
type
init
prop
seq
type
y
yhat
ei

Iteration 6

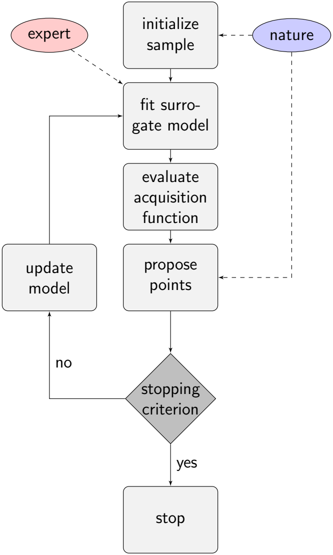

Iter = 7, Gap = 3.8848e−01
25
20
15
10
5
0
0.06
0.04
0.02
0.00
−20
0
20
x
y
ei
type
init
prop
seq
type
y
yhat
ei

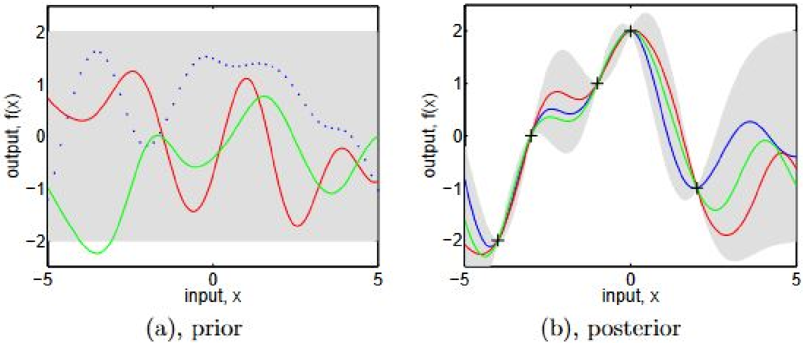
Iteration 7



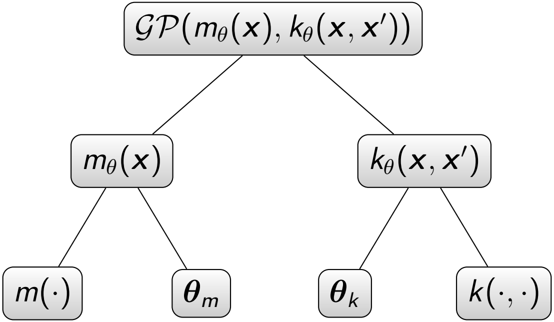
Iteration 10

Bayesian Optimization



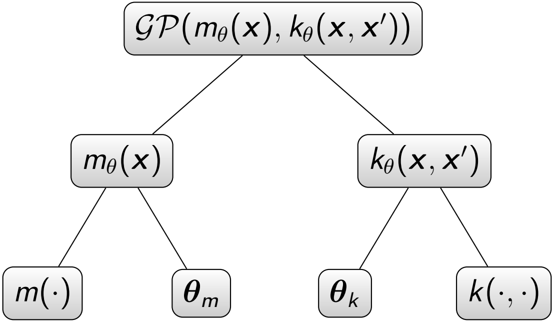


Functional GP regression: Three functions drawn from prior (a) and posterior (b) GP. Image credits: [Rasmussen, 2003].



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.

Agenda

- **Mean parameters** influence convergence the most, followed by the **kernel's functional form**.
- **Mean functional form** and **Kernel parameters** play a (relatively) negligible role.

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

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

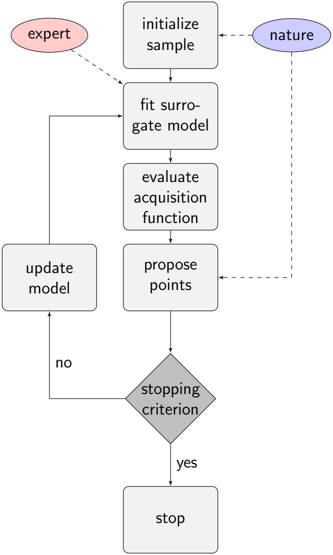
given a base kernel $k_{\boldsymbol{\theta}}(x, x')$ and a degree of imprecision $c > 0$.

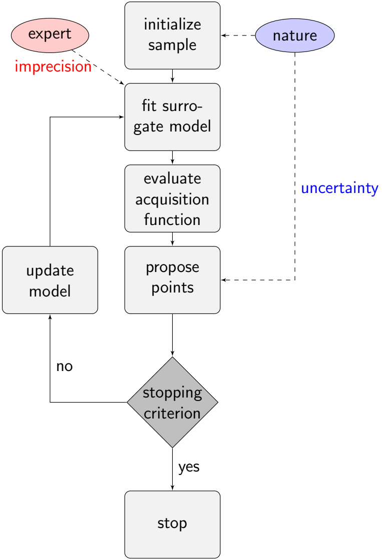
$\rightarrow$ results in a set of posteriors whose upper and lower mean estimates $\underline{\hat{\mu}}(x)_c$, $\overline{\hat{\mu}}(x)_c$ can be derived

Note: See [Benavoli and Zaffalon, 2015] for more on imprecise Gaussian processes and prior near-ignorance.

Generalized Lower Confidence Bound (GLCB)

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


expert
initialize sample
nature
fit surrogate model
evaluate acquisition function
update model
propose points
no
stopping criterion
yes
stop


initialize sample
expert
imprecision
nature
fit surro- gate model
evaluate acquisition function
uncertainty
update model
propose points
no
stopping criterion
yes
stop

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 $\boldsymbol{k}_x = [k_{\boldsymbol{\theta}}(x, x_1), ..., k_{\boldsymbol{\theta}}(x, x_n)]^T$, the base kernel matrix $\boldsymbol{K}_n$ (from training) and the degree of imprecision c. For some[1] values of c (depending on observations):

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

[1] For a thorough case distinction, please refer to the Appendix.

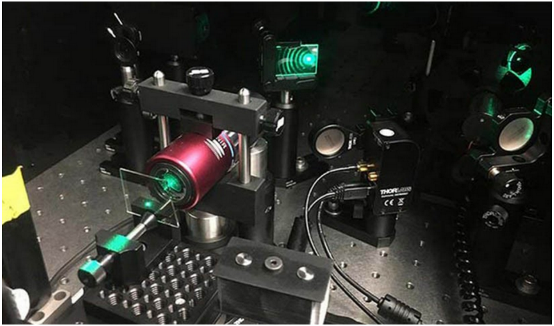
For sufficiently high c, the model imprecision $\overline{\hat{\mu}}(\boldsymbol{x}) - \underline{\hat{\mu}}(\boldsymbol{x})$ even simplifies further:

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

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

LMU

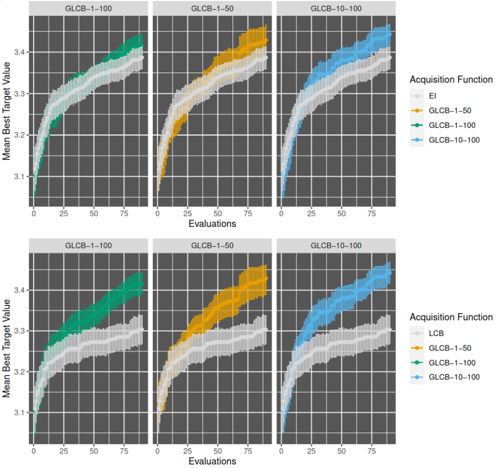
Application in Material Science



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

GLCB – Results



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
    - multivariate extensions
    - 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!

- 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`

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.

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

Literature II

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

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.

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

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

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.

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:

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

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. Comparisons between mean and kernel are more valid than between functional form and parameters.

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 $x \in \mathcal{X}$:

$$\boldsymbol{K}_n = [k_\theta(x_i, x_j')]_{ij}. \tag{3}$$

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 $\boldsymbol{S}_k = 1_n^T \boldsymbol{K}_n^{-1} 1_n$.

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

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

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

Upper and lower mean estimates

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

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

$$\underline{\hat{\mu}}(x) = \boldsymbol{k}_x^T \boldsymbol{K}_n^{-1} \boldsymbol{y} + (1 - \boldsymbol{k}_x^T \boldsymbol{s}_k) \frac{\boldsymbol{s}_k^T \boldsymbol{y}}{c + \boldsymbol{S}_k} \tag{7}$$