# Advancing Bayesian Optimization: The Mixed-Global-Local (MGL) Kernel and Length-Scale Cool Down


**Kim Peter Wabersich, Marc Toussaint — NIPS BayesOpt Workshop, December 2016**

Machine Learning & Robotics Lab, University of Stuttgart, wabersich@kimpeter.de, marc.toussaint@informatik.uni-stuttgart.de



**Abstract:** Bayesian Optimization (BO) has become a core method for solving expensive black-box optimization problems. While much research focussed on the choice of the acquisition function, we focus on online length-scale adaption and the choice of kernel function. Instead of choosing hyperparameters in view of maximum likelihood on past data, we propose to use the acquisition function to decide on hyperparameter adaptation more robustly and in view of the future optimization progress. Further, we propose a particular kernel function that includes non-stationarity and local anisotropy and thereby implicitly integrates the efficiency of local convex optimization with global Bayesian optimization. Comparisons to state-of-the art BO methods underline the efficiency of these mechanisms on global optimization benchmarks.


## Ideas & Related Work

### Lengthscale Cool-Down based on acquisition function (AR)

- Choose hyperprior to accelerate optimization $\rightarrow$ more acquisition
- Neglect model fit up to a *best case* correlation lower bound

### Mixed-Global-Local (MGL) Kernel

- A novel kernel function to represent local convex polynomial regions
- Implies optimization steps analogous to classical (quasi-Newton-type) model-based optimization combined with global Bayesian optimization

### Related Work


- Ziyu Wang, et al (2016) Bayesian optimization in a billion dimensions via random embeddings, Journal of Artificial Intelligence Research
- Hossein Mohammadi, et al (2016) Small ensembles of kriging models for optimization, arXiv preprint arXiv:1603.02638
- Ruben Martinez-Cantin (2015) Locally-Biased Bayesian Optimization using Nonstationary Gaussian Processes, NIPS workshop on Bayesian Optimization


### General Bayesian Optimization

1. Given an initial set of samples $\{X_1, \boldsymbol{y}_1\}$, prior $\mathcal{GP}(c_\mu, k)$ and acquisition function $\alpha$
2. iterate $n = 1$ until $N$:
3. perform model adaption with $\{X_n, \boldsymbol{y}_n\}$
4. $\boldsymbol{x}_n = \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{D}} \alpha_n(\boldsymbol{x})$ and extend set $\{X_1, \boldsymbol{y}_1\}$ by evaluation of objective function at $\boldsymbol{x}_n$
5. return best observation

## Length-Scale Cool Down

### Choosing the hyperprior to accelerate optimization

- Online length-scale cool down method based on the acquisition function instead of model selection, like e.g. using maximum-likelihood
- Let

$$\alpha_{r,n} := \frac{\alpha^*(\tilde{l}_n)}{\alpha^*(l_{n-1})} \tag{1}$$

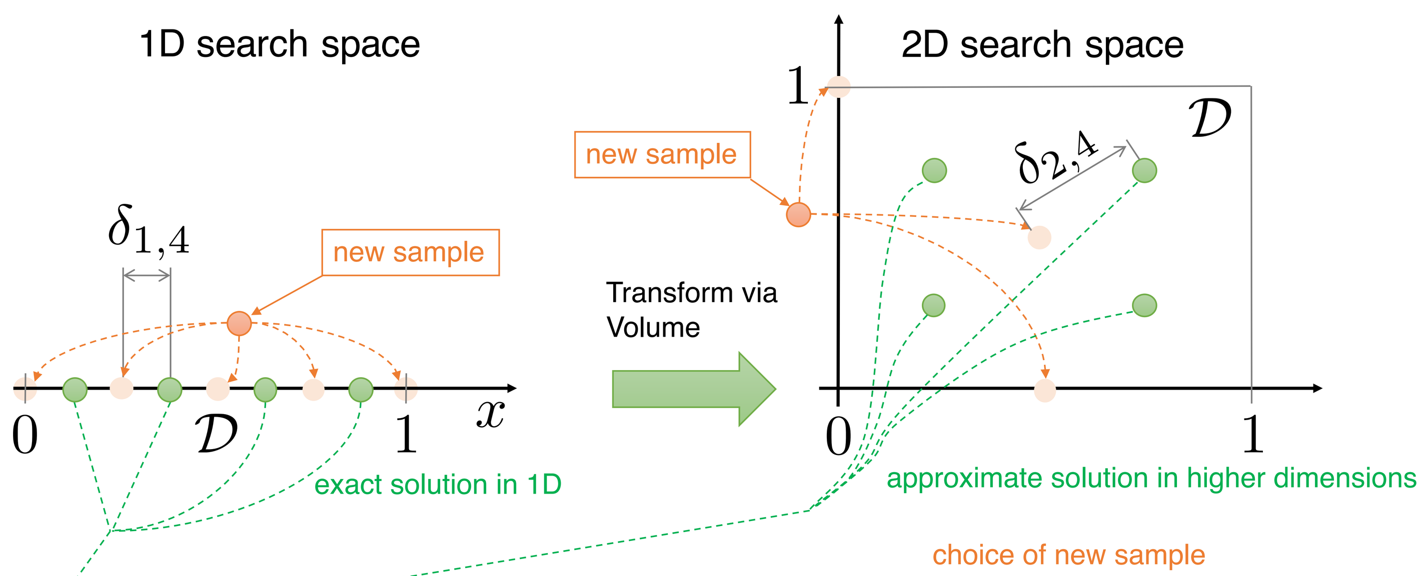
be the alpha-ratio, where $\alpha^*(l) = \min_{\boldsymbol{x}\in\mathcal{D}} \alpha_n(\boldsymbol{x}; l)$ is the optimal aquisition with length-scale $l$ and $\tilde{l}_n < l_{n-1}$ is a smaller *candidate* length-scale

- Typically a *smaller* length-scale leads to *larger* variance $\Rightarrow \alpha_{r,n} > 1$

$\Downarrow$ *Turn this argument around*

If $\alpha_{r,n}$ is *not substantially* larger than 1, decreasing the length-scale will typically *not* yield better chances for progress in the optimization

- Lower bound based on minimal correlation for "best case" set $\bar{X}_n$:

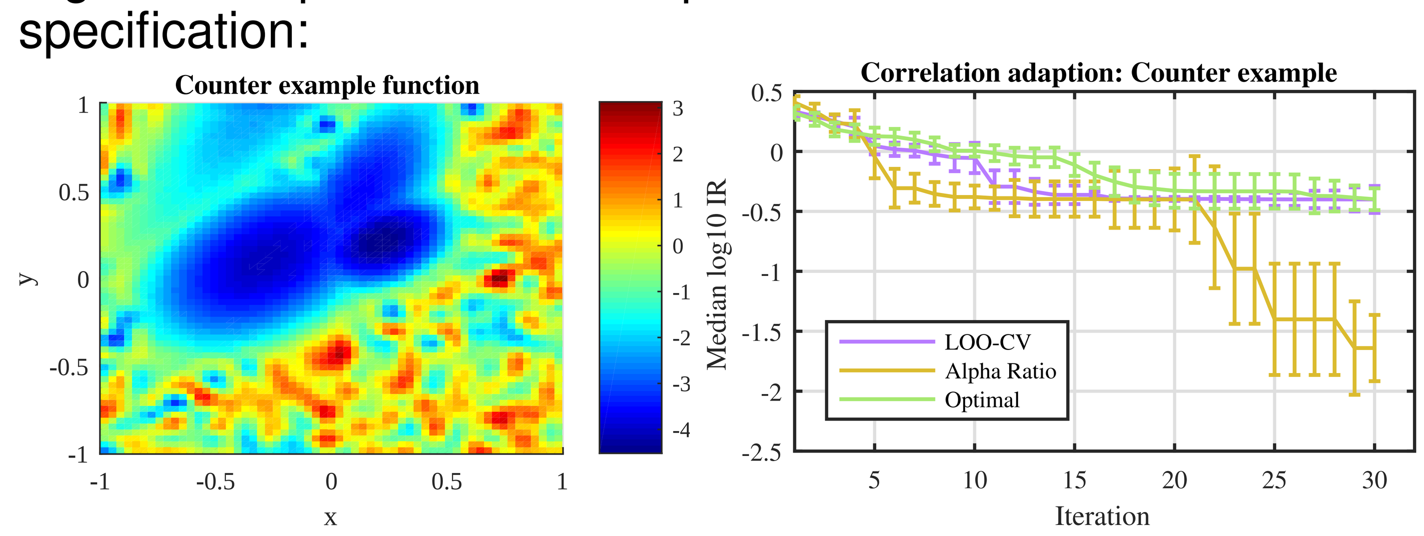



$$\bar{X}_n := \operatorname{argmax}_{X,\ \text{s.t.}\ |X|=n, x'\in X} \left(\min_{x\in\mathcal{D}} k(||x - x'||; l)\right)$$

With $\delta_{d,n}$ and a desired best case correlation $\bar{c}$, we get for the Squared Exponential kernel:

$$\bar{l}_n(d, \bar{c}) = \sqrt{-\frac{1}{2\log(\bar{c})}} \left( \frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{3}{2})} \pi^{0.5(1-d)} \underbrace{\frac{1}{n}}_{\delta_{1,n}} \right)^{\frac{1}{d}} \tag{2}$$

- Pseudo code for adjusting length-scale
  1. calculate lower bound $\bar{l}_n(d, \bar{c})$ (Eq. 2)
  2. choose $\tilde{l}_n \leftarrow \max\{l_{n-1}/2, \bar{l}_n(d, \bar{c})\}$
  3. $\alpha^*(l_{n-1}) \leftarrow \min_{\boldsymbol{x}\in\mathcal{D}} \alpha_n(\boldsymbol{x}; l_{n-1})$ acquisition with current length-scale $l_{n-1}$
  4. $\alpha^*(\tilde{l}_n) \leftarrow \min_{\boldsymbol{x}\in\mathcal{D}} \alpha_n(\boldsymbol{x}; \tilde{l}_n)$ acquisition with $\tilde{l}_n$
  5. $\alpha_{r,n} \leftarrow \alpha^*(\tilde{l}_n)/\alpha^*(l_{n-1})$
  6. based on a threshold on $\alpha_{r,n}$ keep lengthscale or reduce to $\tilde{l}_n$
- Significant performance improvements in case of model miss-specification:




## Mixed-Global-Local (MGL) Kernel

### Formalize the intuition: How to model a (local) minimum?

- Given a data set $D = \{(\boldsymbol{x}_i, y_i)\}$, we call a convex subset $\mathcal{U} \subset \mathcal{D}$ a convex neighborhood if the solution of the regression problem

$$\{\beta_0^*, \boldsymbol{\beta}_1^*, B^*\} = \operatorname*{argmin}_{\beta_0, \boldsymbol{\beta}_1, B} \sum_{k:\boldsymbol{x}_k\in\mathcal{U}} \left[ (\beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{x}_k + \frac{1}{2}\boldsymbol{x}_k^T B \boldsymbol{x}_k) - y_k \right]^2$$

($\boldsymbol{x}_k \in \mathcal{U}$ the data points in $\mathcal{U}$) has a positive definite Hessian $B$

- The Mixed-Global-Local (MGL) kernel is given by

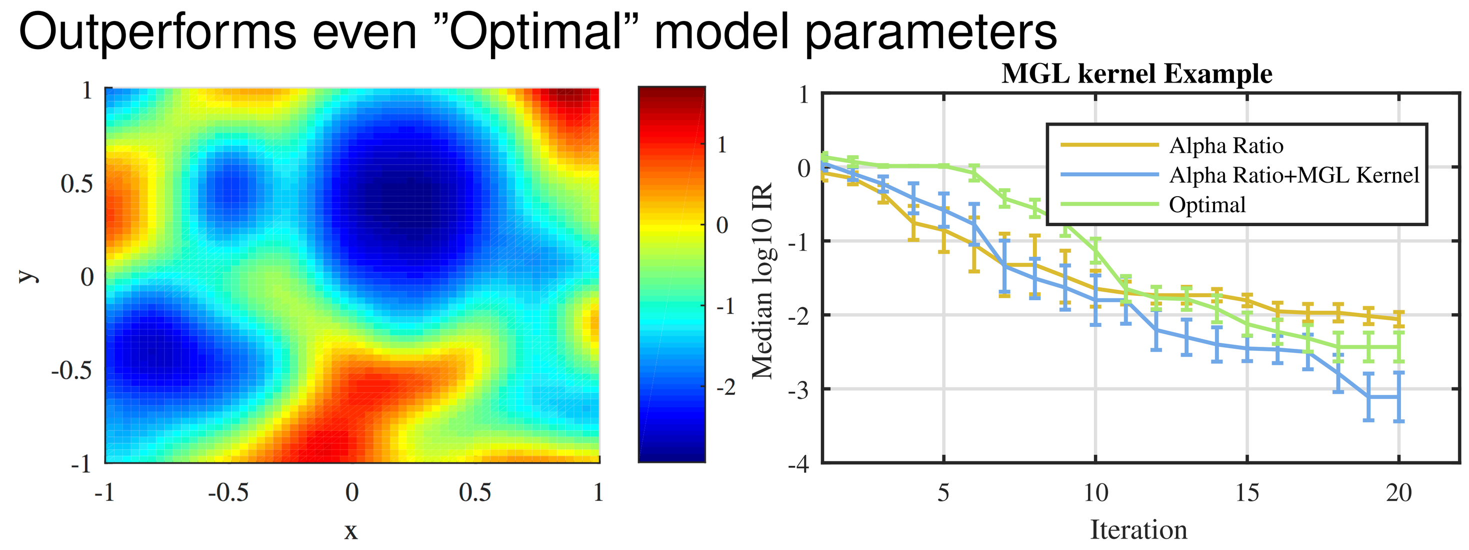
$$k_{\text{MGL}}(\boldsymbol{x}, \boldsymbol{x}') = \begin{cases} k_q(\boldsymbol{x}, \boldsymbol{x}'),\ \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{U}_i, \\ k_s(\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{x} \notin \mathcal{U}_i, \boldsymbol{x}' \notin \mathcal{U}_j \\ 0,\ \text{else} \end{cases}$$

for any $i, j$, where $k_s$ is a stationary-isotropic kernel and

$$k_q(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T \boldsymbol{x}' + 1)^2$$

the quadratic kernel

- Construct $\mathcal{U}_i$ by KNN-search: Start at each sample point and gradually increase K, check KNN for qualifying as $\mathcal{U}_i$ candidate. Choose best $\mathcal{U}_i$'s
- Outperforms even "Optimal" model parameters




## Results

- Results for combined length-scale cool down based on alpha ratio and MGL kernel (**AR+MGL**) vs. Predictive Entropy Search (**PES**), Infinite Metric GP Optimization (**IMGPO**), and Expected Improvement (**EI**) with 'optimal' chosen hyperparameters
- Median of 32 runs, variance estimate via Bootstrapping
- **Software** and **extended paper version** can be found at `www.kimpeter.de`